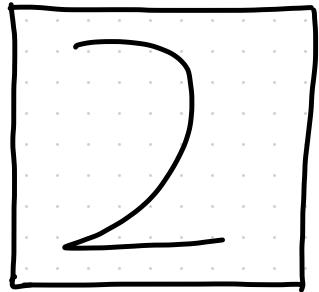


RECENT SUCCESSES OF NN

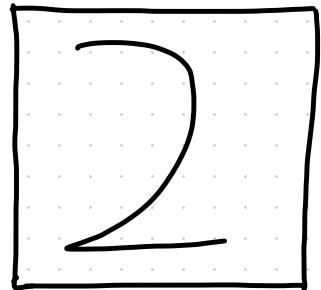
- * State-of-the-art (SOTA) in most fields

PARADIGM CHANGE

PARADIGM CHANGE

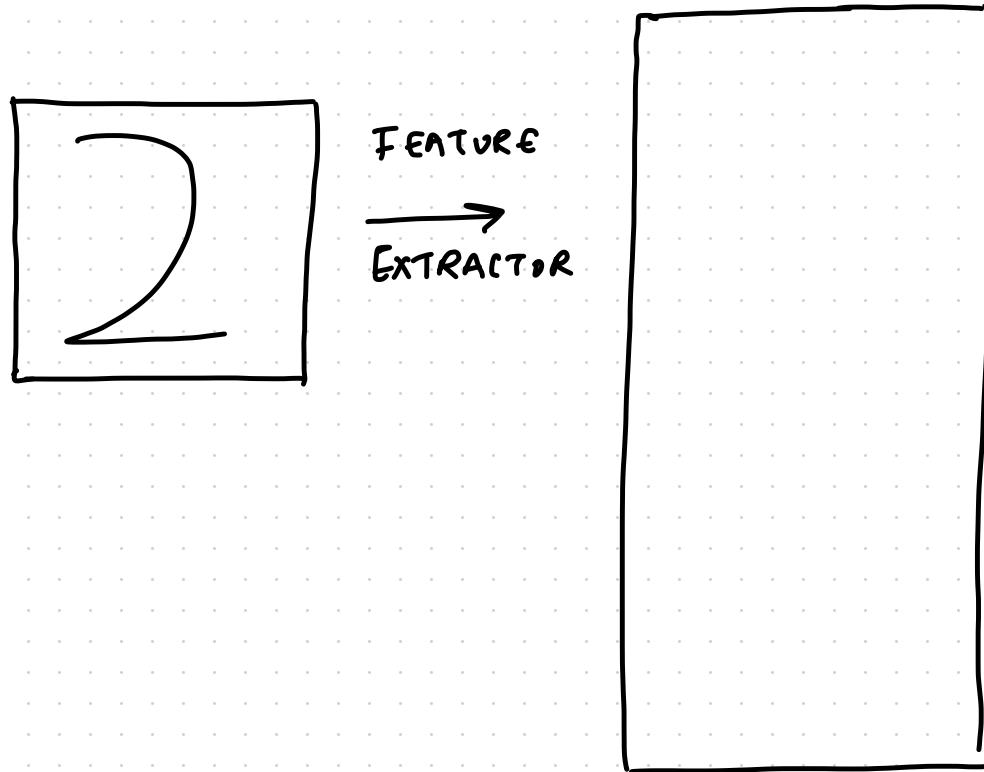


PARADIGM CHANGE

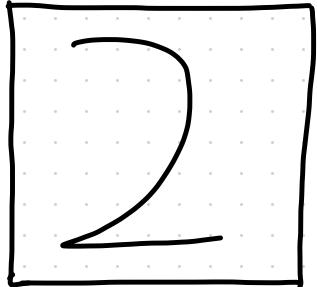


FEATURE
→
EXTRACTOR

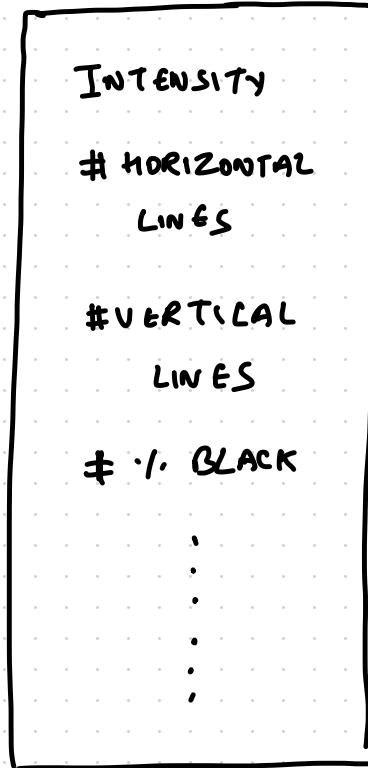
PARADIGM CHANGE



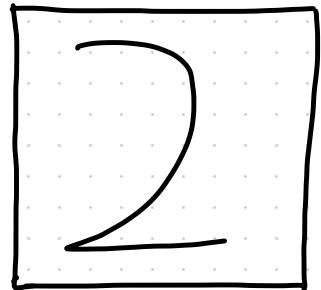
PARADIGM CHANGE



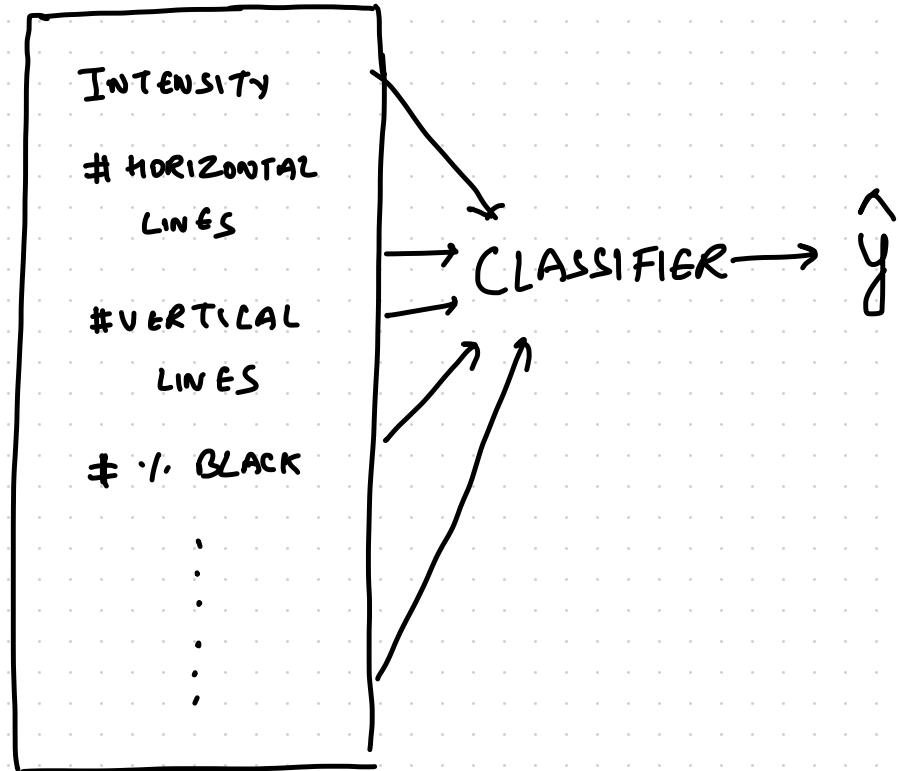
FEATURE
EXTRACTOR



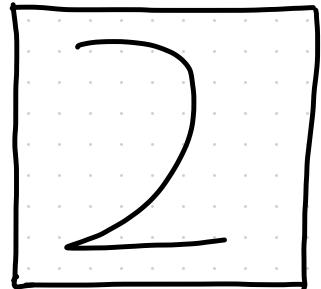
PARADIGM CHANGE



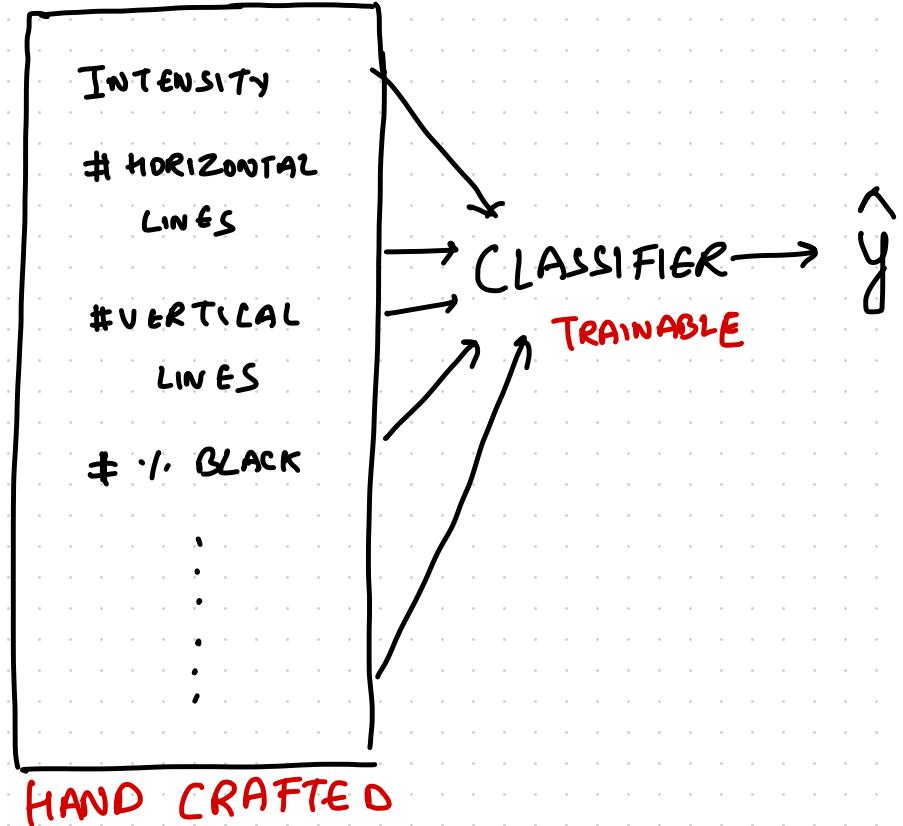
FEATURE
EXTRACTOR



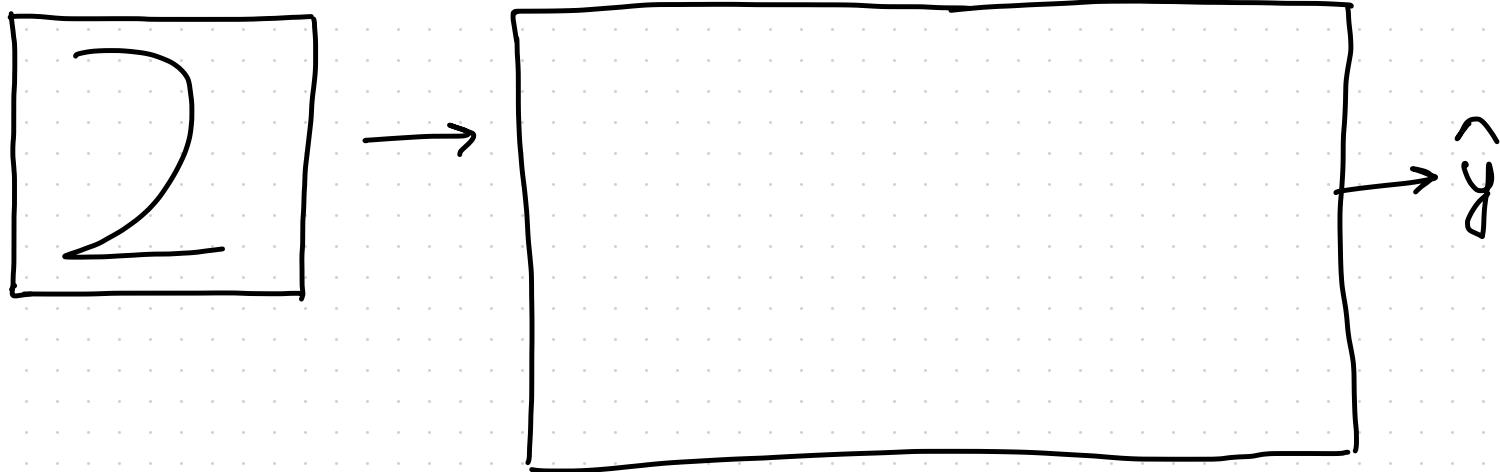
PARADIGM CHANGE



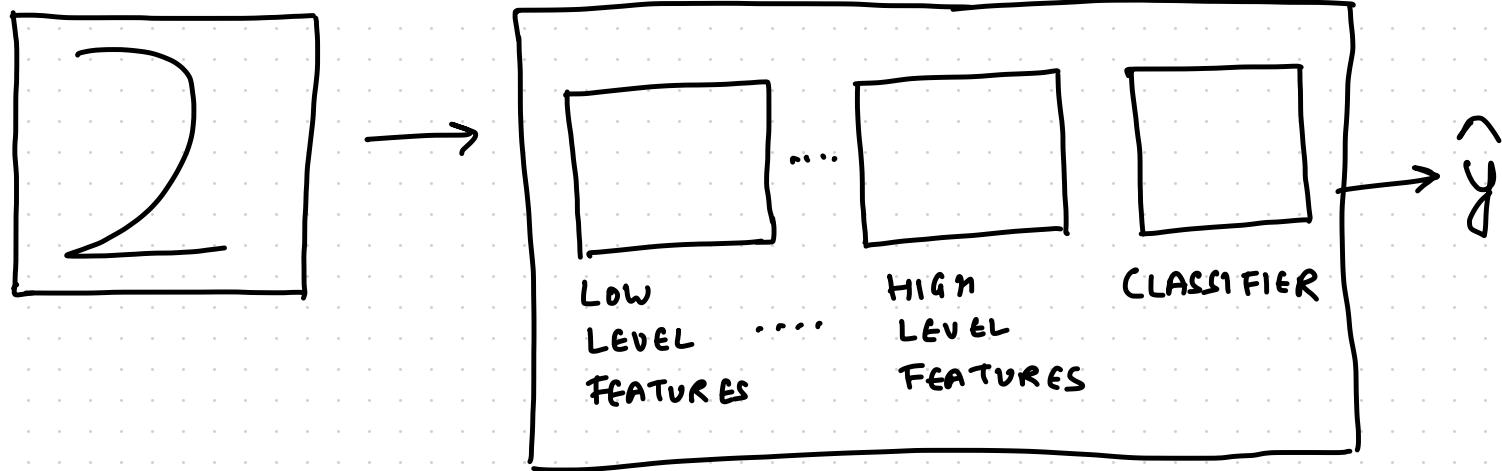
FEATURE
EXTRACTOR



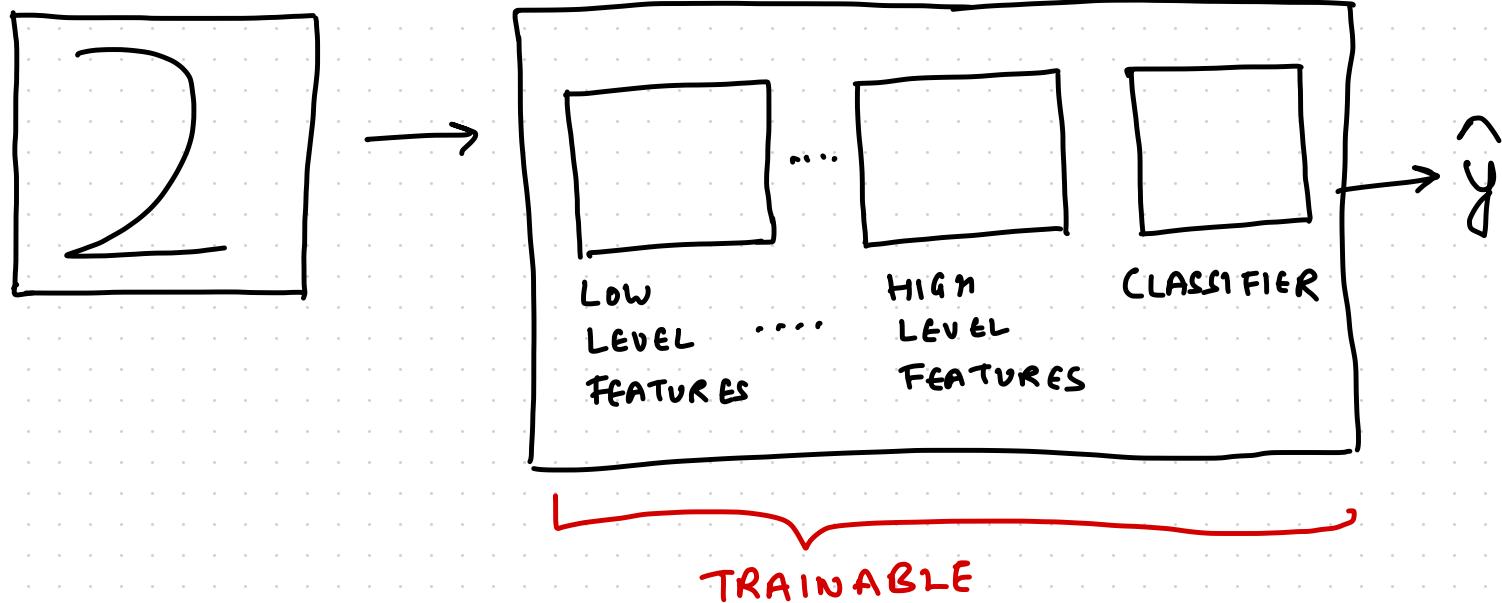
PARADIGM CHANGE (NNS)



PARADIGM CHANGE (NNs)



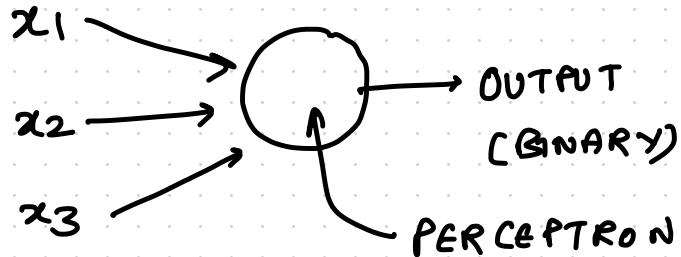
PARADIGM CHANGE (NNs)



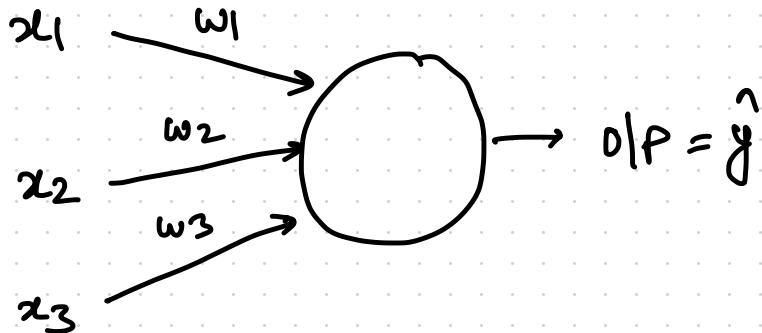
PERCEPTRON

— ARTIFICIAL NEURON DEVELOPED BY
ROSENBLATT IN 1960^s INSPIRED BY
MC CULLOCH & PITTS

BINARY IP

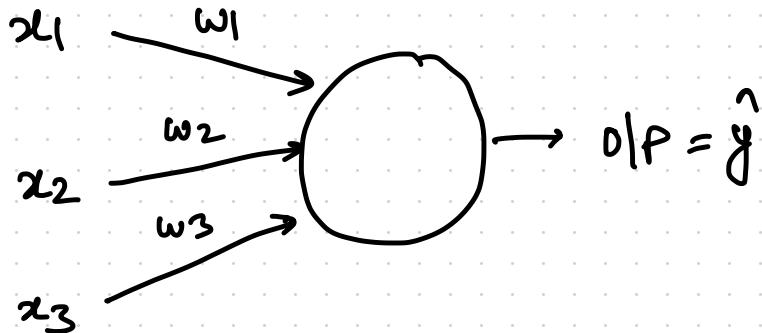


PERCEPTRON



$$O/P = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i \leq \text{THRESHOLD} \\ 1 & ; \sum w_i x_i > \text{THRESHOLD} \end{cases}$$

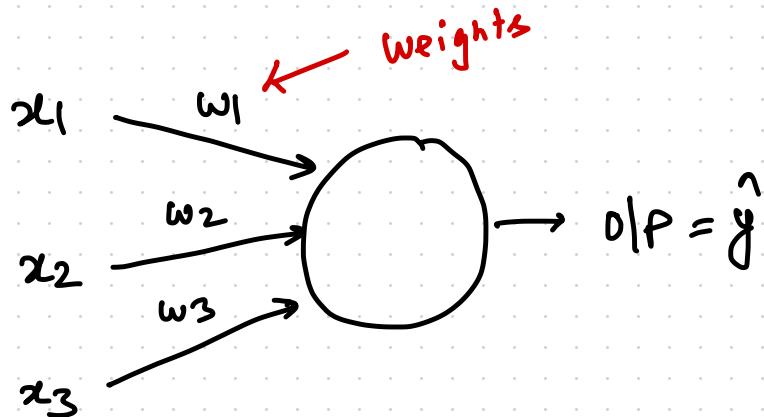
PERCEPTRON



$$O/P = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i \leq \text{THRESHOLD} \\ 1 & ; \sum w_i x_i > \text{THRESHOLD} \end{cases}$$

NEURONS "FIRE" ABOVE THRESHOLD

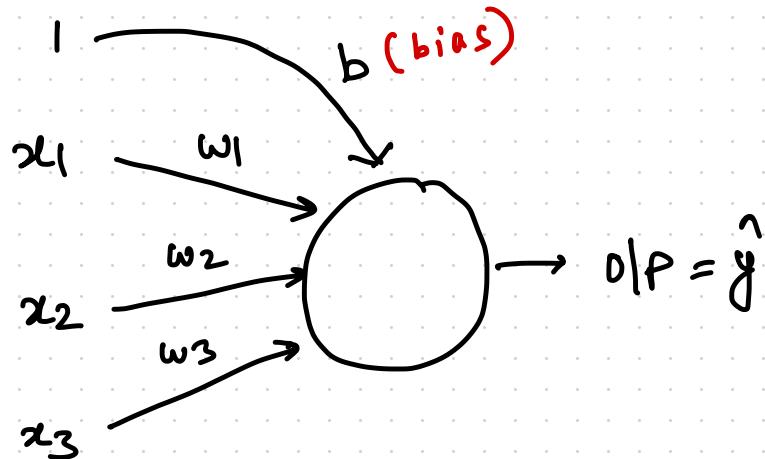
PERCEPTRON



$$O/P = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i \leq \text{THRESHOLD} \\ 1 & ; \sum w_i x_i > \text{THRESHOLD} \end{cases}$$

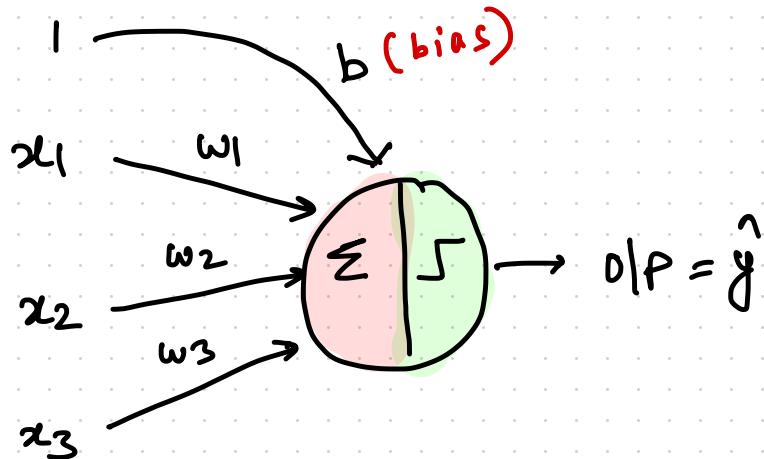
NEURONS "FIRE" ABOVE THRESHOLD

PERCEPTRON



$$O|P = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i + b \leq 0 \\ 1 & ; \sum w_i x_i + b > 0 \end{cases}$$

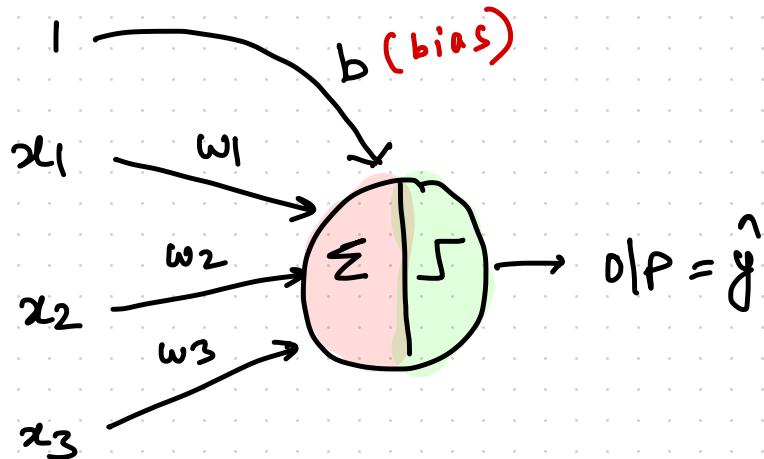
PERCEPTRON



NEURON HAS 2 COMPONENTS

- ① SUMMATION
- ② ACTIVATION : STEP FUNCTION

PERCEPTRON



NEURON HAS 2 COMPONENTS

① SUMMATION

② ACTIVATION : STEP FUNCTION

(SIGMOID)
Activation



LEARNING BINARY GATES

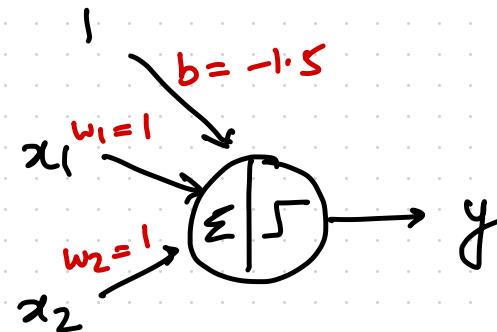
Q) FOR 2 IIPS x_1 & x_2 learn w's and b for
BINARY AND

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

LEARNING BINARY GATES

Q) FOR 2 IIPS x_1 & x_2 learn w_i 's and b for
BINARY AND

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1



LEARNING BINARY GATES

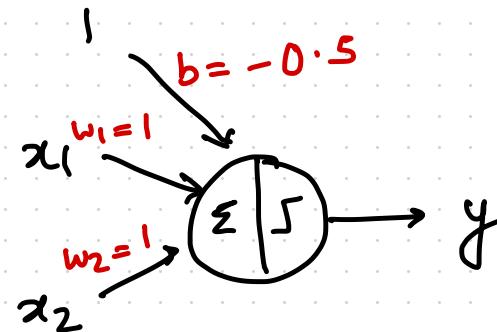
Q) FOR 2 IIPS x_1 & x_2 learn w's and b for
BINARY OR

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	1

LEARNING BINARY GATES

Q) FOR 2 IIPS x_1 & x_2 learn w's and b for
BINARY OR

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	1



LEARNING BINARY GATES

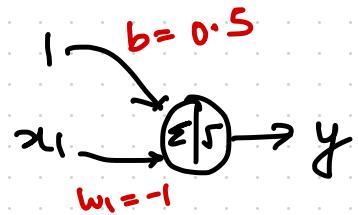
Q) FOR 1 IIPS x_1 learn w's and b for
UNARY NOT

x_1		y
0		1
1		0

LEARNING BINARY GATES

Q) FOR 1 IIPS x_1 learn w_i 's and b for UNARY NOT

x_1	y
0	1
1	0



LEARNING BINARY GATES

Q) FOR 2 IIPS $x_1 \& x_2$ learn w's and b for
NAND

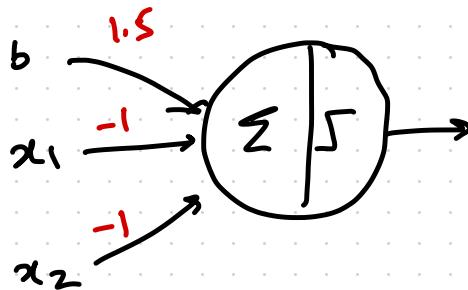
x_1	x_2	y
0	0	1
0	1	1
1	0	1
1	1	0

LEARNING BINARY GATES

Q) FOR 2 IIPS $x_1 \& x_2$ learn w's and b for
NAND

x_1	x_2	y
0	0	1
0	1	1
1	0	1
1	1	0

APPROACH #1

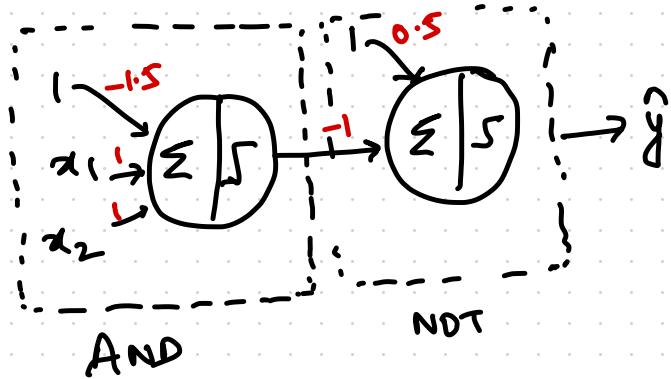
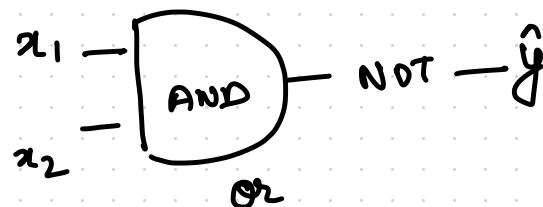


LEARNING BINARY GATES

Q) FOR 2 IIPS $x_1 \& x_2$ learn w's and b for
NAND

APPROACH #2

x_1	x_2	y
0	0	1
0	1	1
1	0	1
1	1	0



PERCEPTRON LEARNING ALGORITHM

IP: $X \rightarrow_{N \times D} ; y \rightarrow_{N \times 1} ; lr \rightarrow \text{learning rate} ; it \rightarrow \# \text{iterations}$

PERCEPTRON LEARNING ALGORITHM

IP: $X \rightarrow_{N \times D} Y \rightarrow_{N \times 1}$; $lr \rightarrow$ learning rate; $it \rightarrow$ #iterations

SI AUGMENT $x \rightarrow x' = s \cdot x + \begin{bmatrix} 1 \\ \vdots \end{bmatrix}$; $x'[1:s] = x$; $x'[s] = 1$

PERCEPTRON LEARNING ALGORITHM

IP: $X \xrightarrow[N \times D]{} y \xrightarrow[N \times 1]{} ; lr \xrightarrow{\text{learning rate}} ; it \xrightarrow{\# \text{iterations}}$

S1 AUGMENT $x \rightarrow x' - s.t. x'[1:s] = x ; x'[s] = [1;]$

S2 INIT $w \in \mathbb{R}^{D+1}$

PERCEPTRON LEARNING ALGORITHM

IP: $X \rightarrow_{N \times D} Y \rightarrow_{N \times 1}$; $lr \rightarrow$ learning rate; $it \rightarrow$ #iterations

S1 AUGMENT $x \rightarrow x'$ s.t. $x'[1:s] = x$; $x'[s] = [1]$

S2 INIT $w \in \mathbb{R}^{D+1}$

S3 FOR I IN IT:

PERCEPTRON LEARNING ALGORITHM

IP: $X \rightarrow \mathbb{R}^{N \times D}$; $y \rightarrow \mathbb{R}^{N \times 1}$; $lr \rightarrow$ learning rate; $it \rightarrow$ #iterations

S1 AUGMENT $x \rightarrow x' = s.t. x'[1:s] = x$; $x'[1] = [1]$

S2 INIT $w \in \mathbb{R}^D$

S3 FOR i IN IT:

S3.1 FOR n in N :

$$\hat{y} = \text{STEP}(x' \cdot w)$$

$$ERR = y - \hat{y}$$

$$w \leftarrow w + lr * ERR_n * x'[n]$$

PERCEPTRON LEARNING ALGORITHM

IP: $X \rightarrow_{NxD}$; $y \rightarrow_{Nx1}$; $lr \rightarrow$ learning rate; $it \rightarrow$ #iterations

S1 AUGMENT x to x' → s.t. $x'[1:s] = x$; $x'[s] = [1]$

S2 INIT $w \in \mathbb{R}^{D+1}$

S3 FOR I IN IT:

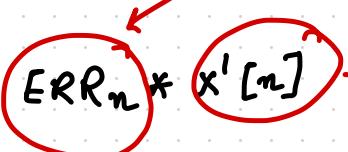
S3.1 FOR n in N :

$$\hat{y} = \text{STEP}(x' \cdot w)$$

$$ERR = y - \hat{y}$$

$$w \leftarrow w + lr * ERR_n * x'[n]$$

ERROR IN n^{th} sample



n^{th} sample

PERCEPTRON LEARNING ALGORITHM

IP: $X \rightarrow_{N \times D} y \rightarrow_{N \times 1}$; lr → learning rate; it → #iterations

S1 AUGMENT x to x' → s.t. $x'[1:s] = x$; $x'[s] = [1; \dots]$

S2 INIT $w \in \mathbb{R}^{D+1}$

S3 FOR I IN IT:

S3.1 FOR n in N :

Analogous to S.G.D

$$\hat{y} = \text{STEP}(x' \cdot w)$$

$$ERR = y - \hat{y}$$

$$w \leftarrow w + lr * ERR_n * x'[n]$$

Analogous to Gradient

ERROR IN n^{th} sample

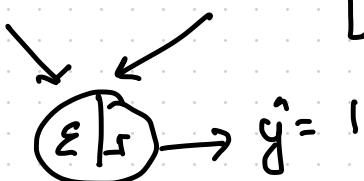
n^{th} sample

PERCEPTRON LEARNING ALGORITHM

$$S31.3 \quad w \leftarrow w + lr * \text{ERR}_n * x'[n]$$

IMAGINE

$$w = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} ; \quad x'[n] = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} ; \quad y = 0$$



PERCEPTRON LEARNING ALGORITHM

S31.3 $w \leftarrow w + lr * \text{ERR}_n * x'[n]$

IMAGINE

$$w = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

$$; \quad x'[n] = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}; \quad y = 0$$

ϵ \rightarrow $\hat{y} = 1$

$$\text{ERR}_n = 0 - 1 = -1$$

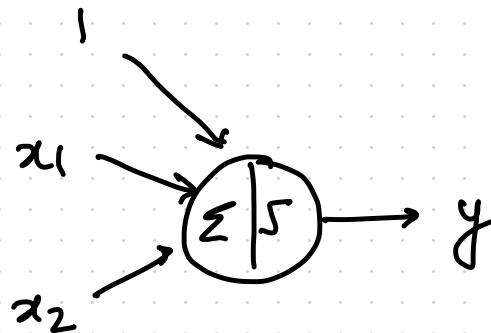
$$w \leftarrow \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} + 0.01 * -1 * \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\leftarrow \begin{bmatrix} 0.99 \\ -1 \\ 1 \end{bmatrix}$$

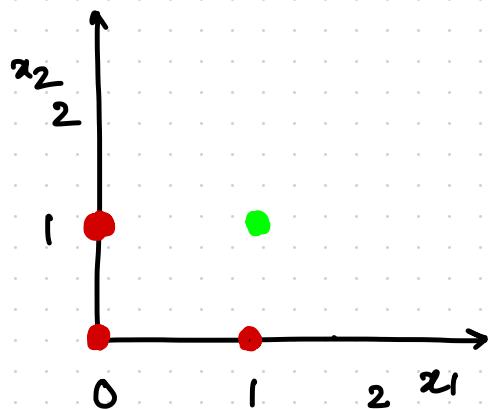
LEARNING BINARY GATES

Q) FOR 2 IPs x_1 & x_2 learn w's and b for
BINARY XOR

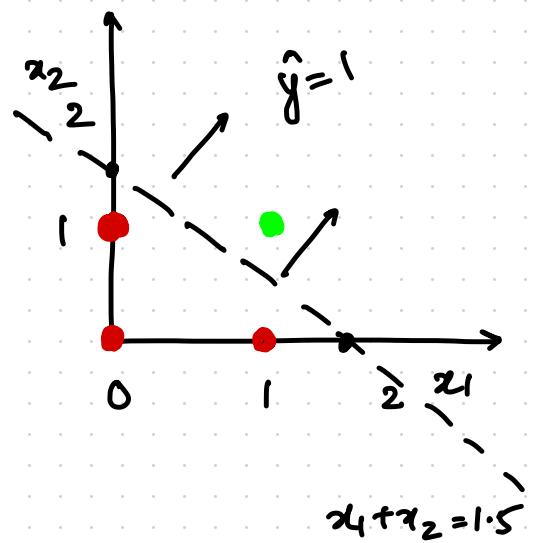
x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0



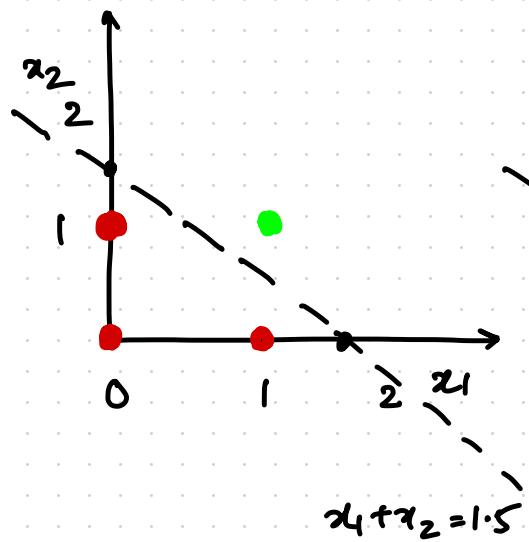
AND



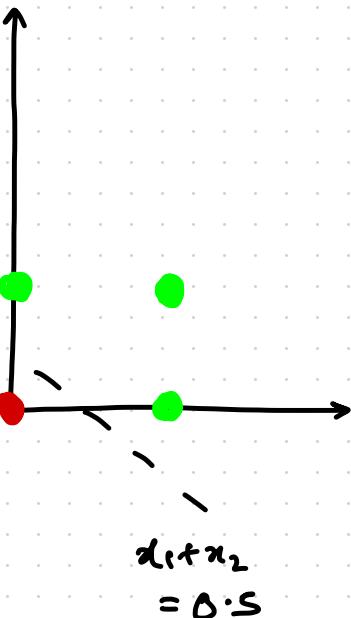
AND



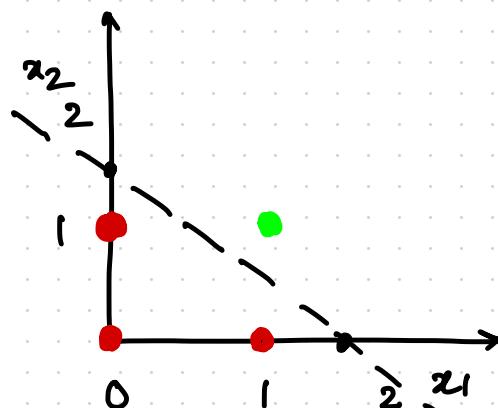
AND



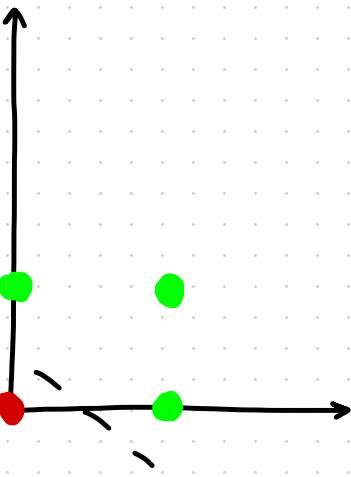
OR



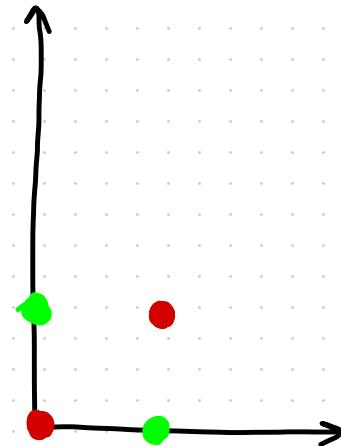
AND



OR



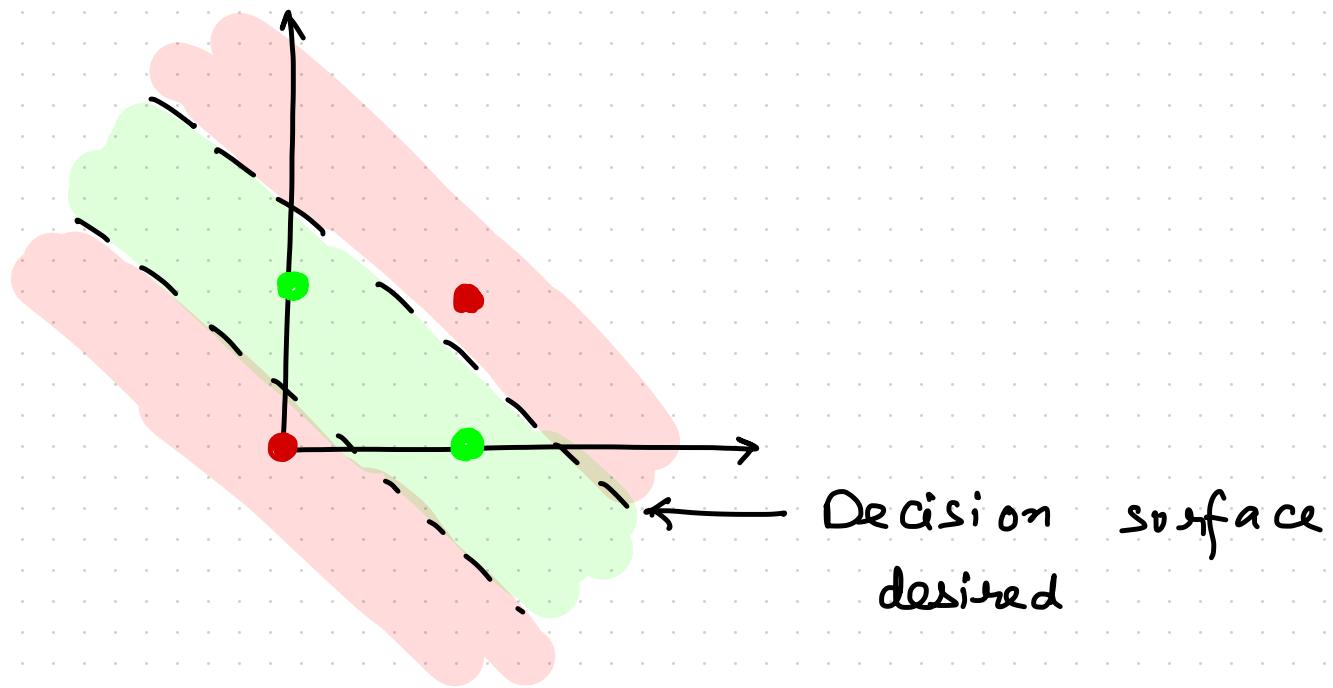
XOR



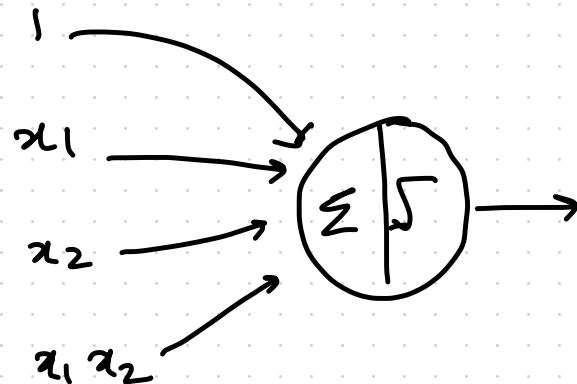
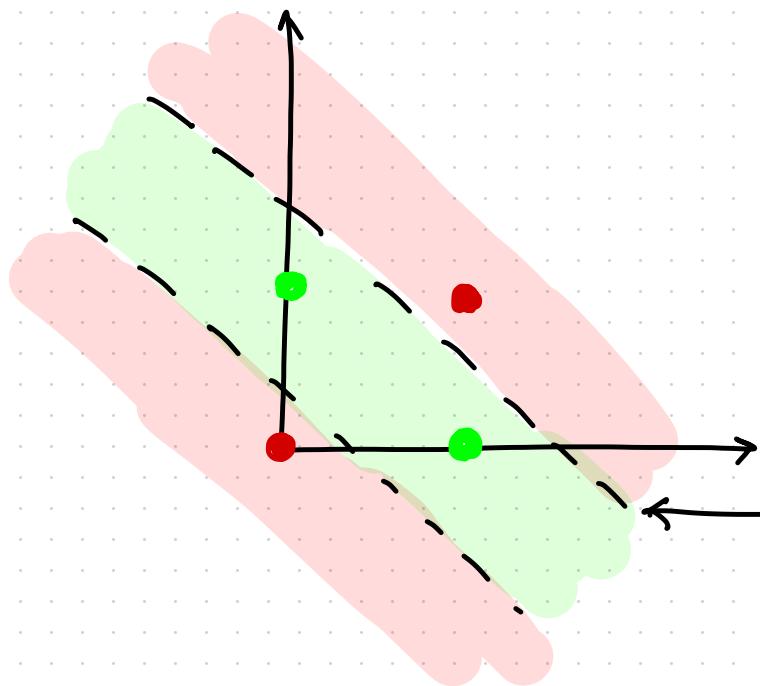
LINEARLY SEPARABLE

NOT
SEPARABLE
LINEARLY

XOR CLASSIFICATION

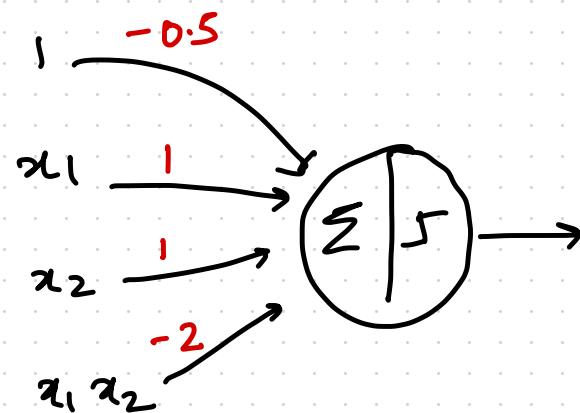
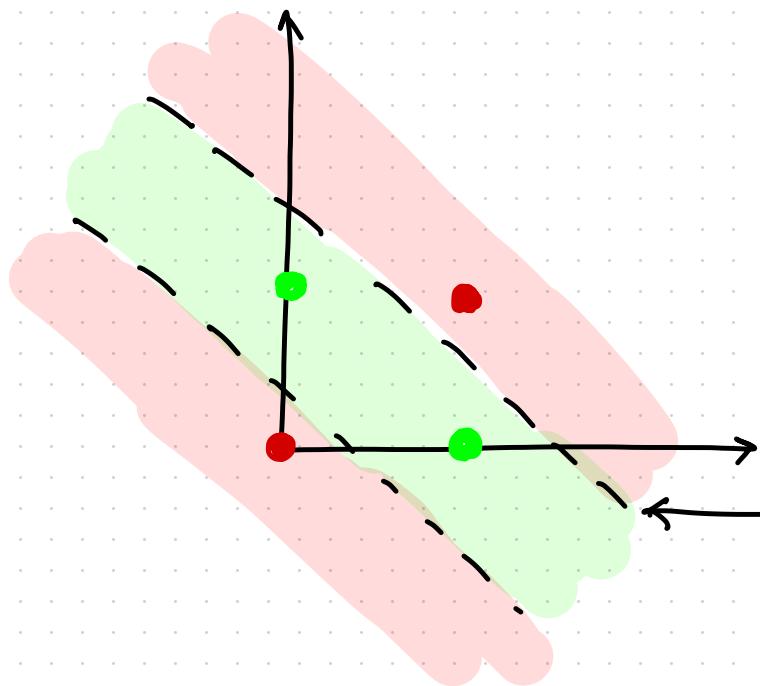


XOR CLASSIFICATION



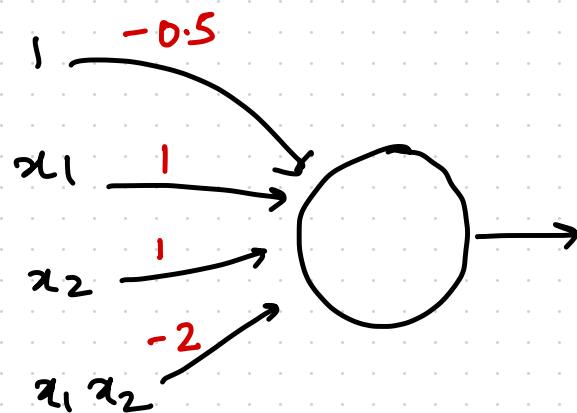
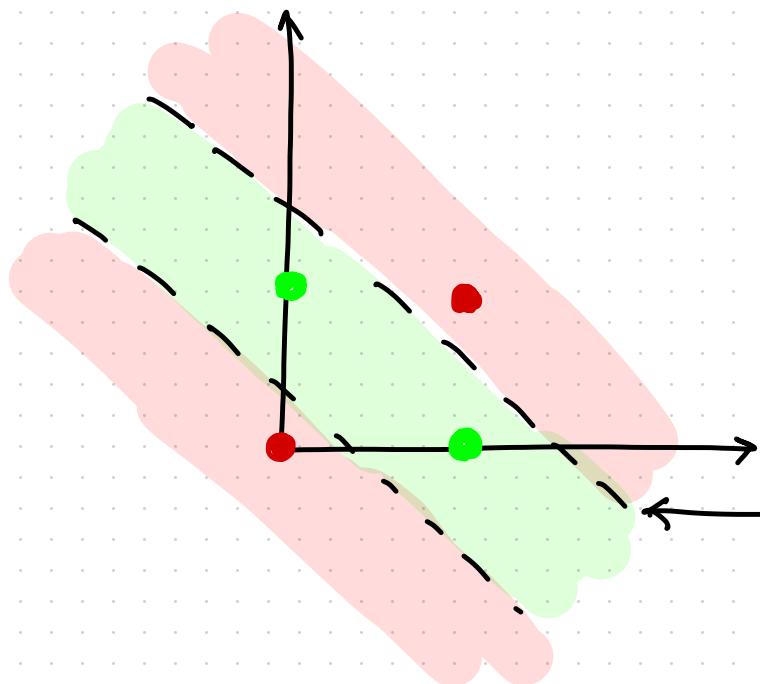
Decision surface
desired

XOR CLASSIFICATION



Decision surface
desired

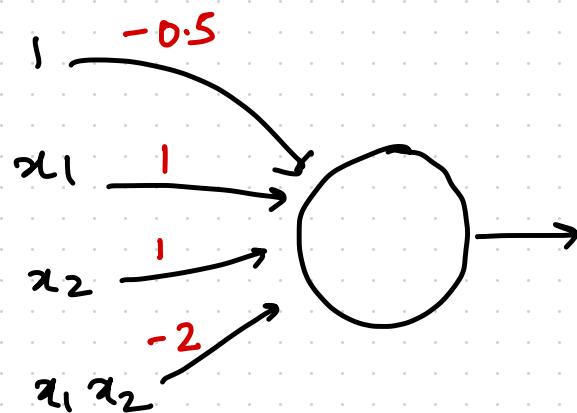
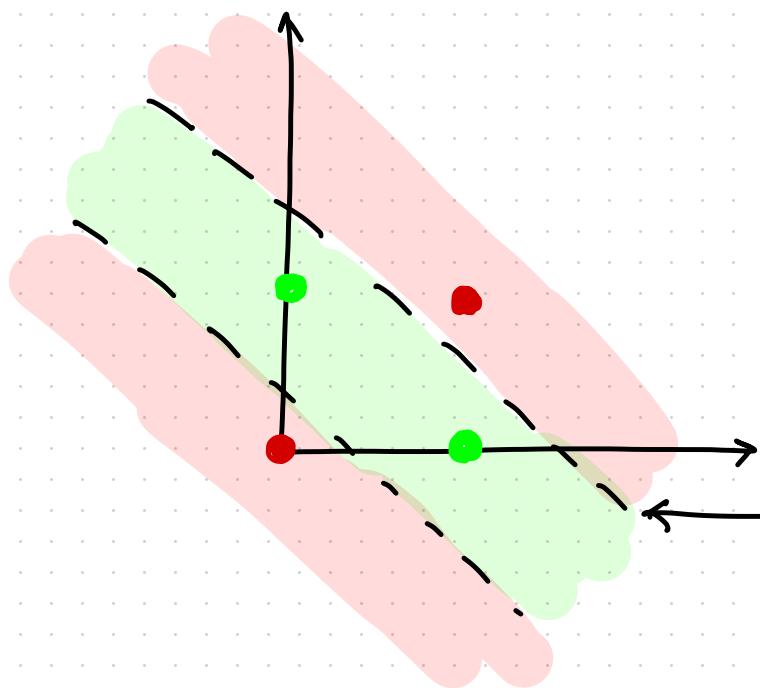
XOR CLASSIFICATION



FOR $x_1 = 0 ; x_2 = 0$; we get

$$\hat{y} = \{-0.5 \leq 0\} = \text{RED CLASS}$$

XOR CLASSIFICATION

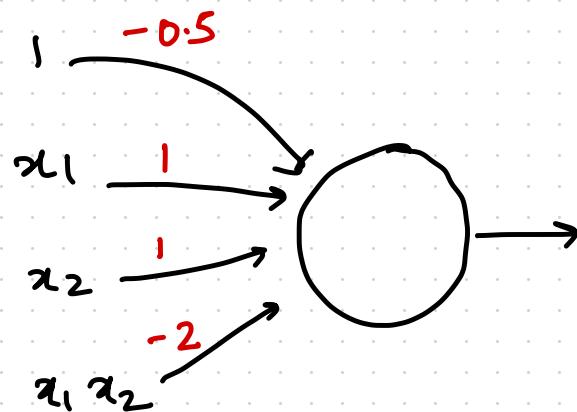
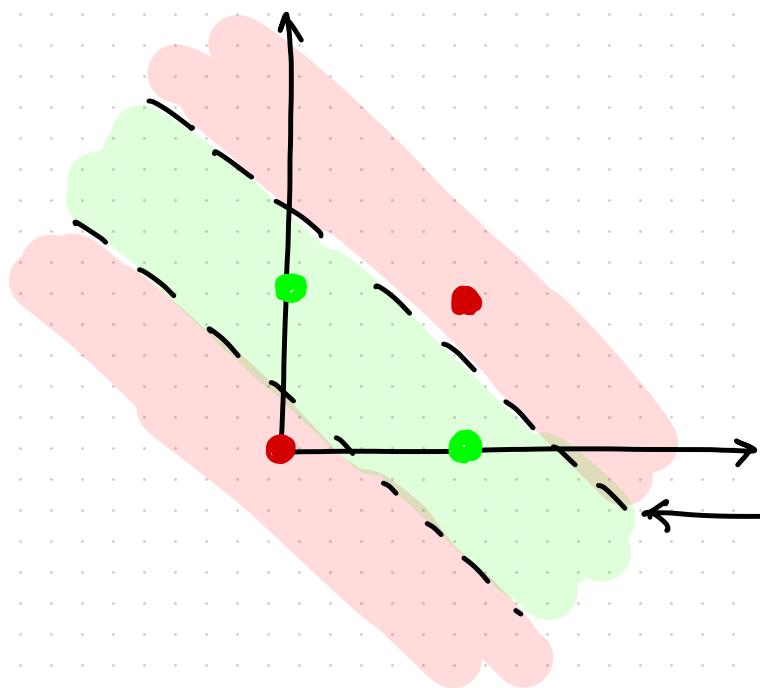


FOR $x_1 = 0 ; x_2 = 1$; we get

$$\hat{y} = \{-0.5 + 1 \leq 0\} = \text{GREEN CLASS}$$

Decision surface
desired

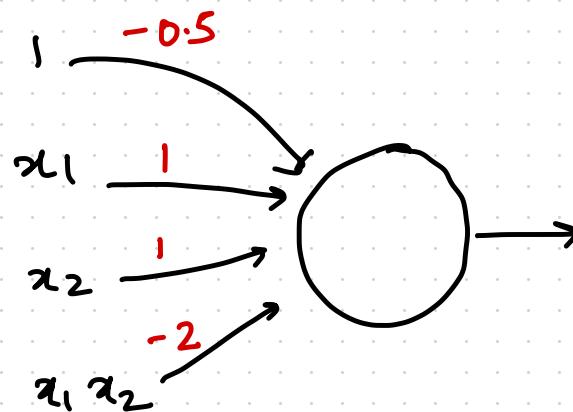
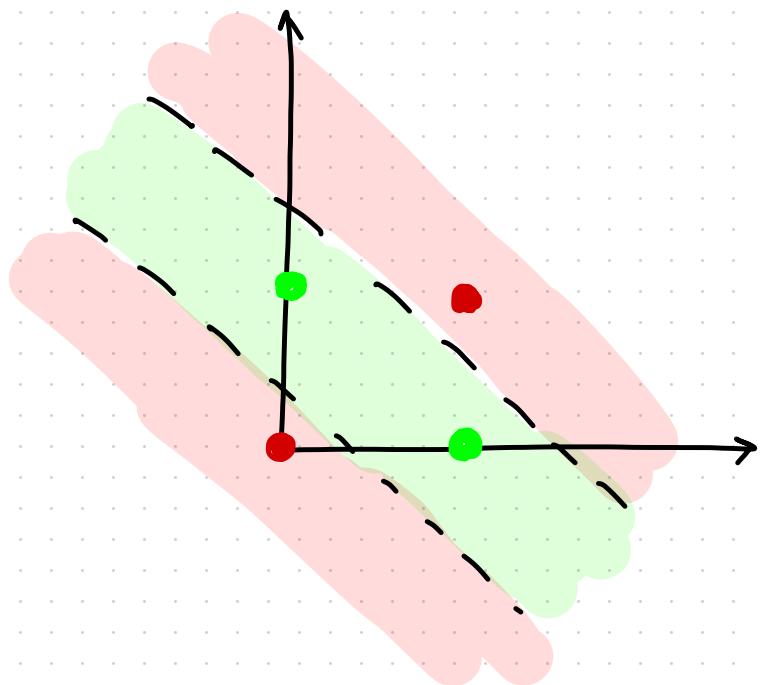
XOR CLASSIFICATION



FOR $x_1 = 1 ; x_2 = 0$; we get
 $\hat{y} = \{ 0.5 \leq 0 \} = \text{GREEN CLASS}$

Decision surface
desired

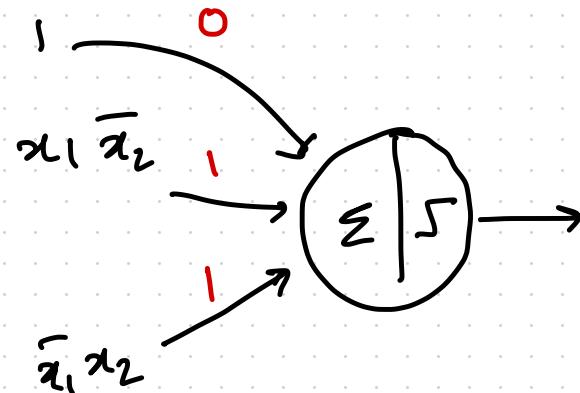
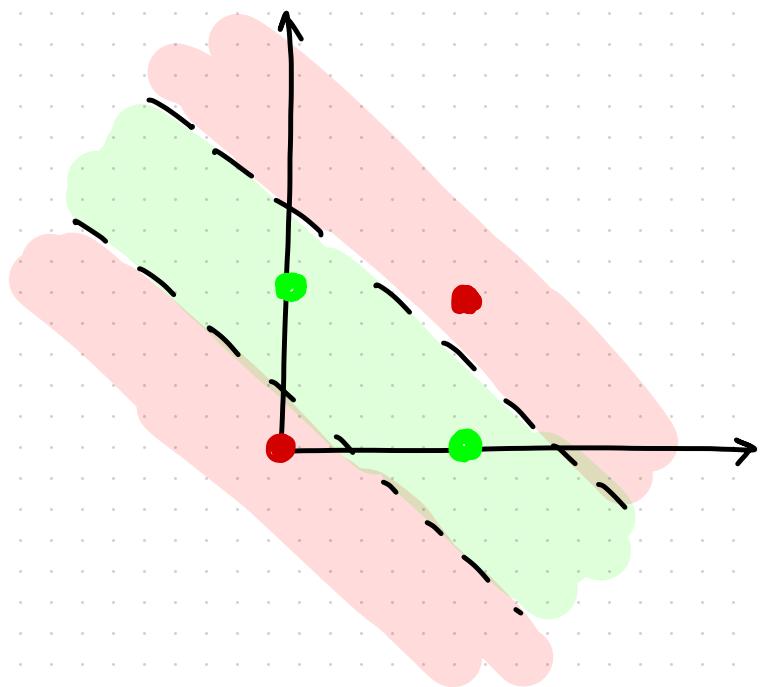
XOR CLASSIFICATION



FOR $x_1 = 1 ; x_2 = 1$; we get
 $\hat{y} = \{-0.5 \leq 0\} = \text{RED CLASS}$

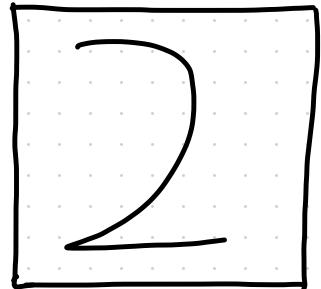
CAN ADD NON-LINEARITY
 BY HAND-CRAFTING
 FEATURES!

XOR CLASSIFICATION

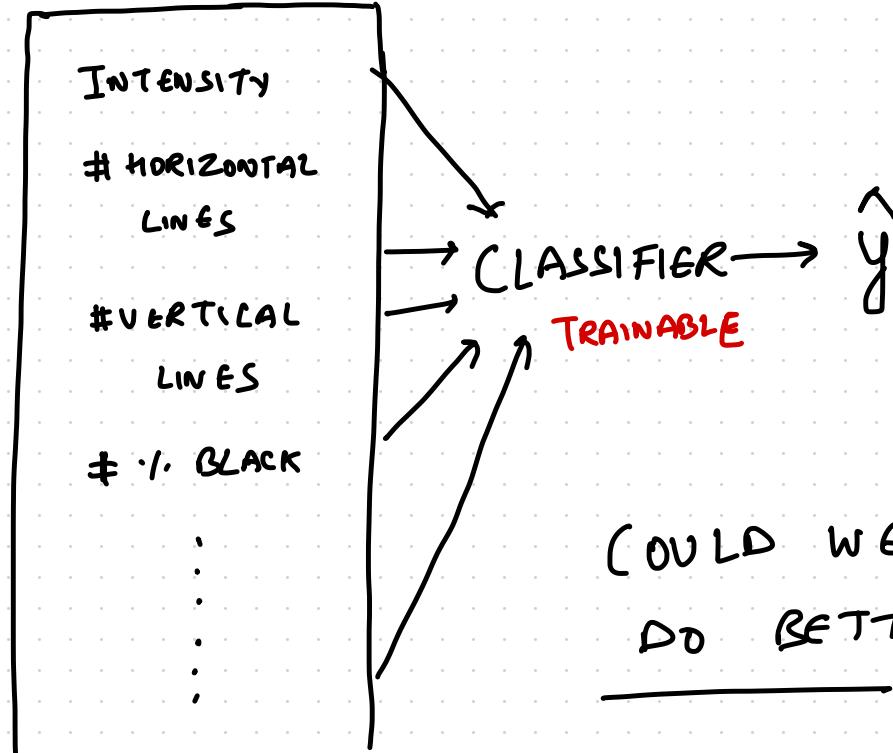


CAN ADD NON-LINEARITY
BY HAND-CRAFTING
FEATURES!

PARADIGM CHANGE



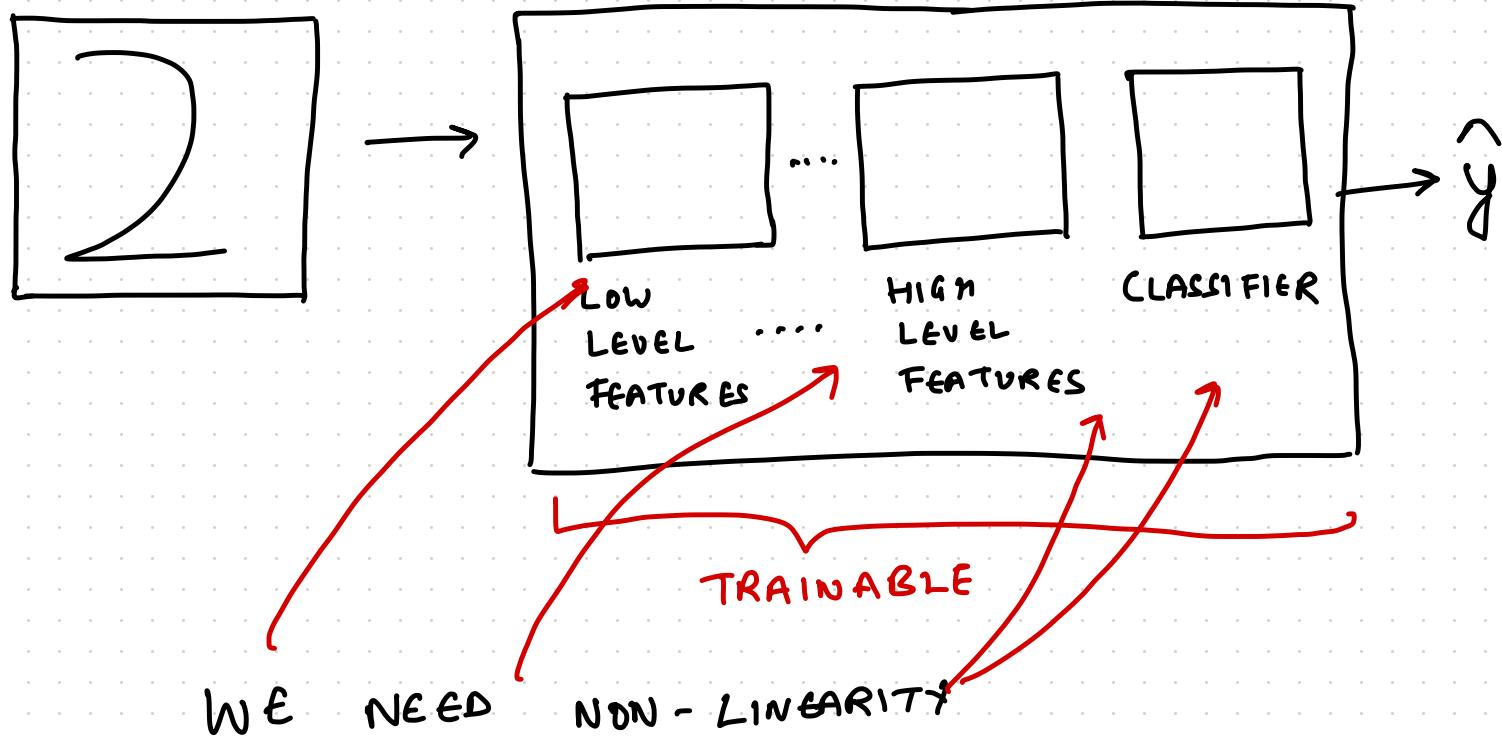
FEATURE
EXTRACTOR



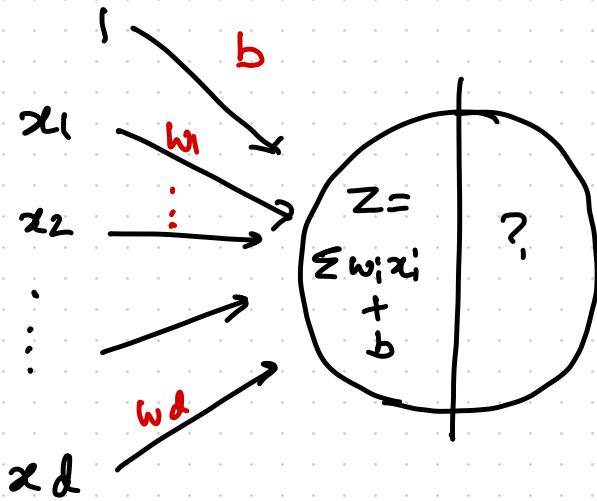
HAND CRAFTED

COULD WE
DO BETTER?

PARADIGM CHANGE (NNs)

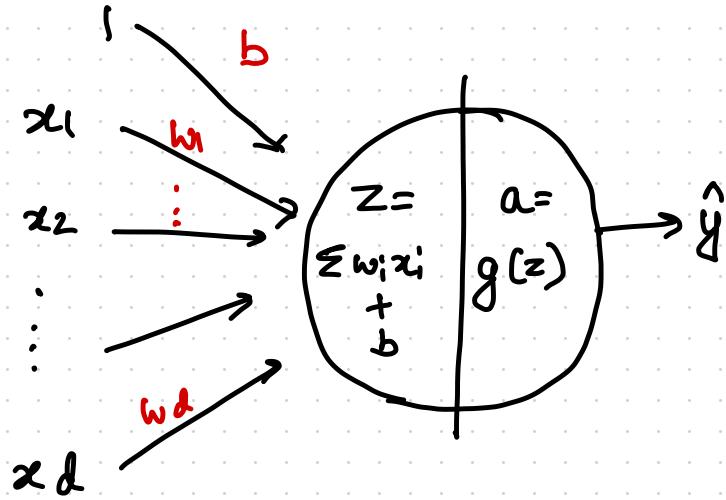


BACK PROPAGATION SUPPORTED ACTIVATIONS



key idea: use
activation
similar to $\begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}$
but differentiable

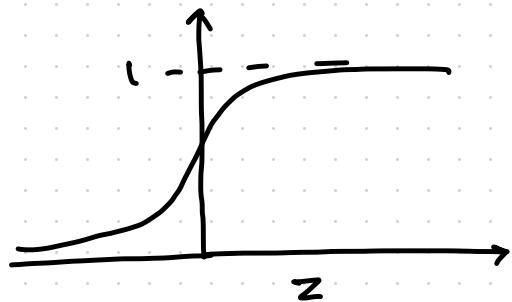
ADDING NON-LINEARITY



$g(z)$: NON-LINEAR
TRANSFORMATION

ACTIVATION FUNCTIONS

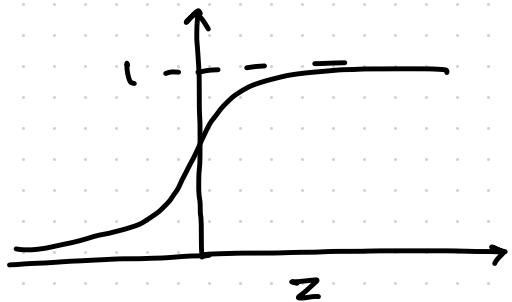
SIGMOID



$$g(z) = \frac{1}{1 + e^{-z}}$$

ACTIVATION FUNCTIONS

SIGMOID



Q): If we have 1 neuron

&

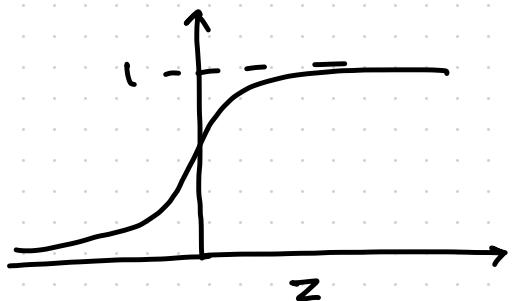
$$g(z) = \frac{1}{1+e^{-z}}$$

what do we
get?

$$g(z) = \frac{1}{1+e^{-z}}$$

ACTIVATION FUNCTIONS

SIGMOID



Q): If we have 1 neuron

&

$$g(z) = \frac{1}{1+e^{-z}}$$

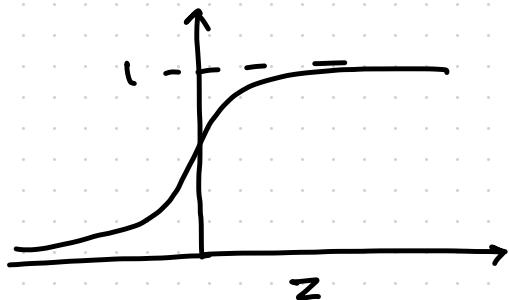
what do we
get?

$$g(z) = \frac{1}{1+e^{-z}}$$

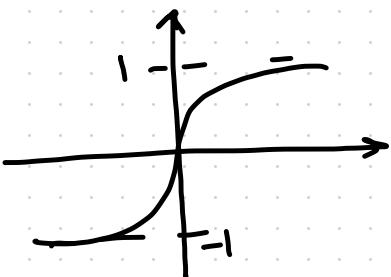
Logistic Regression

ACTIVATION FUNCTIONS

SIGMOID



TANH

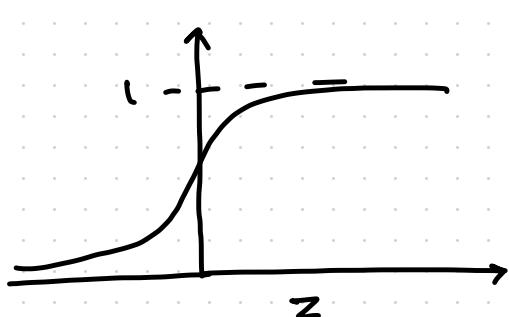


$$g(z) = \frac{1}{1+e^{-z}}$$

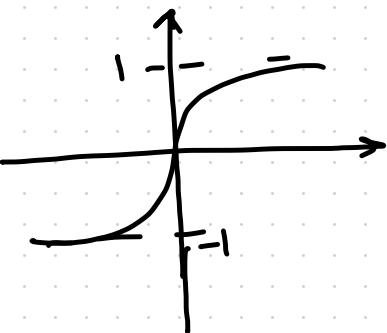
$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

ACTIVATION FUNCTIONS

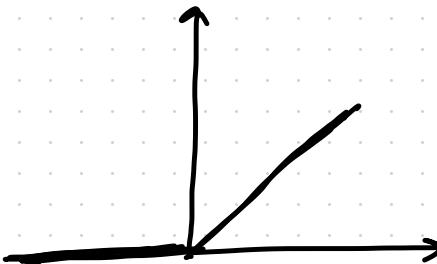
SIGMOID



TANH



ReLU



$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

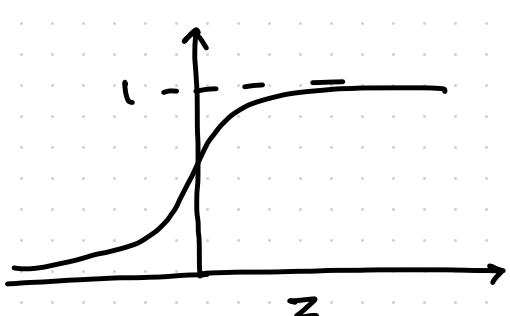
$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & z < 0 \end{cases}$$

or

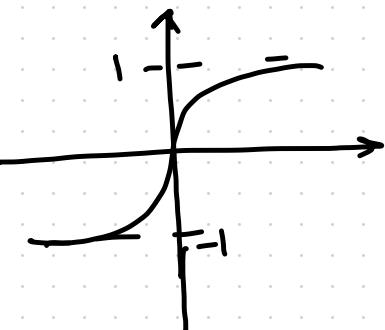
$$g(z) = \max(0, z)$$

ACTIVATION FUNCTIONS

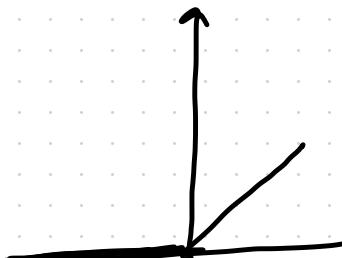
SIGMOID



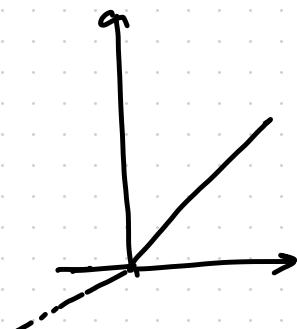
TANH



ReLU



Leaky ReLU



$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & z < 0 \end{cases}$$

or

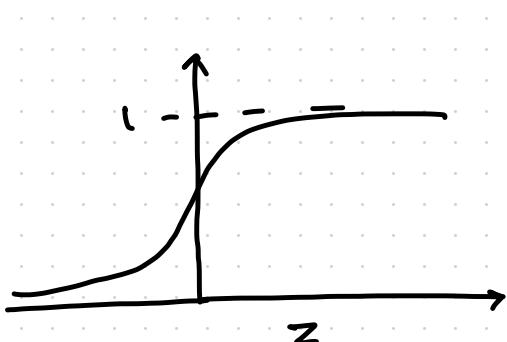
$$g(z) = \max(0, z)$$

$$g(z) = \max(\alpha z, z)$$

$\alpha \rightarrow 0$

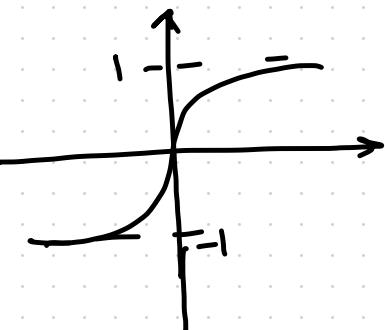
ACTIVATION FUNCTIONS

SIGMOID



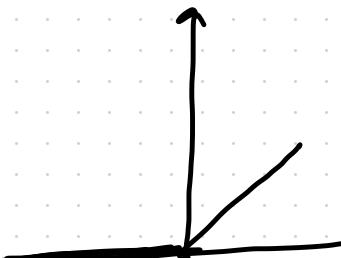
USEFUL FOR
PROBABILISTIC
ESTIMATES
 \therefore Blw 0 & 1

TANH



USEFUL IF
DATA TRANSFORMED
WITH MEAN 0

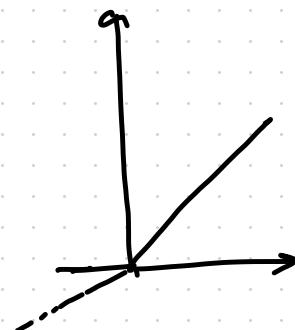
ReLU



GRAD
CHANGER
(Default)

Good
learning
for $|z| = \text{high}$.

Leaky ReLU



Similar

to
ReLU

learns
for $z < 0$
also

DESIRABLE ATTRIBUTES OF ACTIVATION FUNCTIONS

- 1) NON-LINEAR
- 2) (MOSTLY) SMALL CHANGE IN $\text{IF} \Rightarrow$ SMALL
CHANGE IN OP

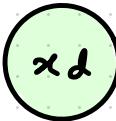
1 LAYER PERCEPTRON (NN)



.

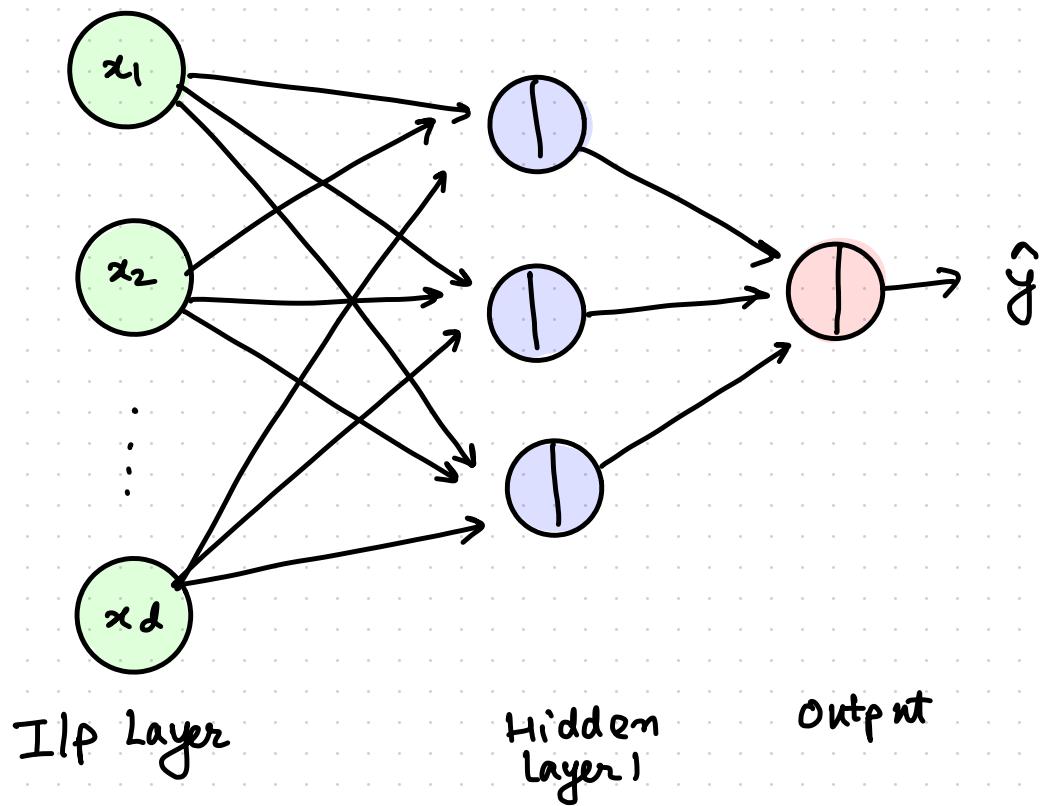
.

.

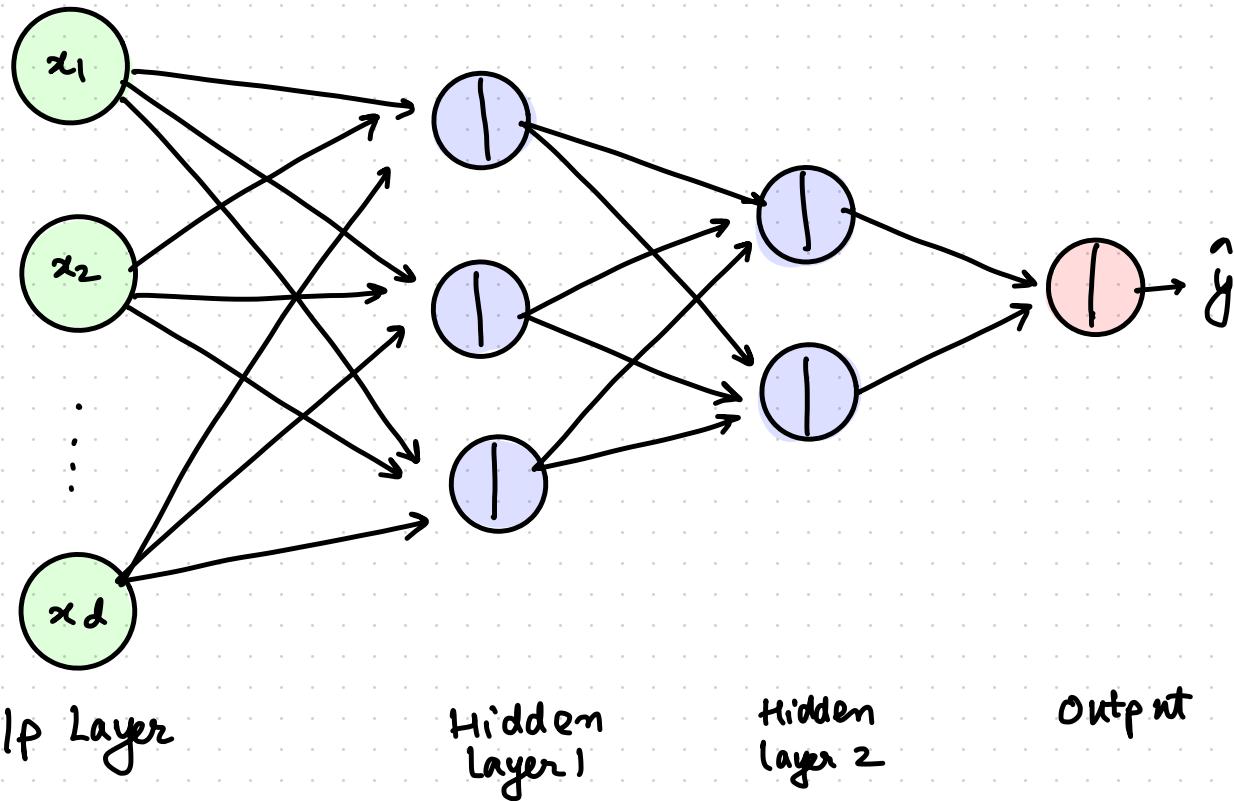


Input Layer

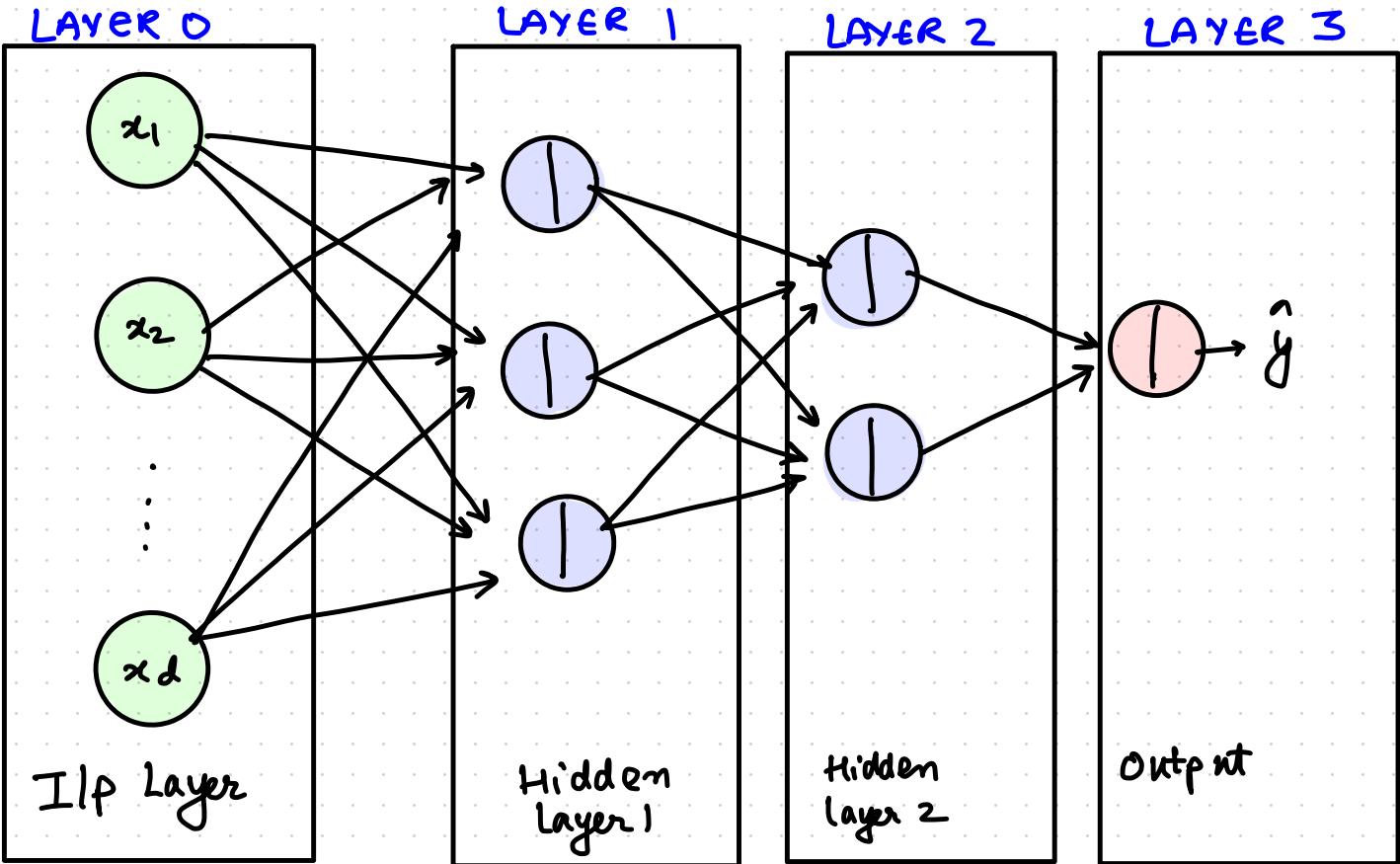
1-LAYER PERCEPTRON (NN)

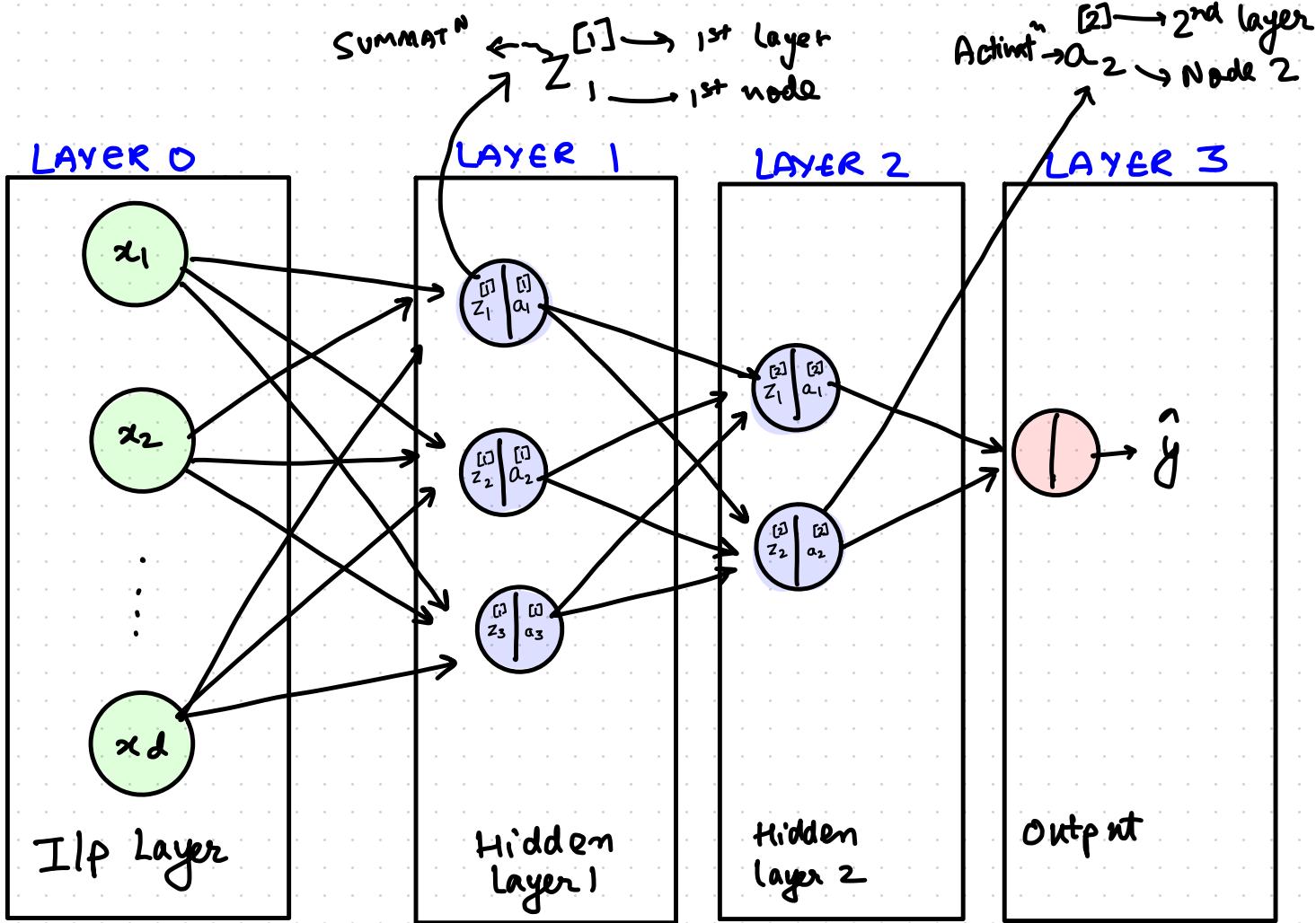


MULTI-LAYER PERCEPTRON

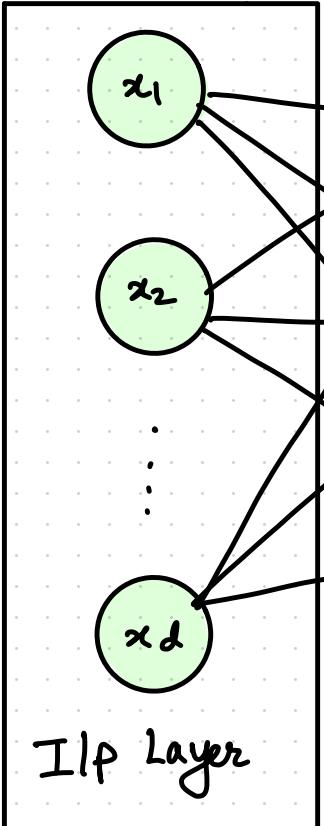


MULTI-LAYER PERCEPTRON

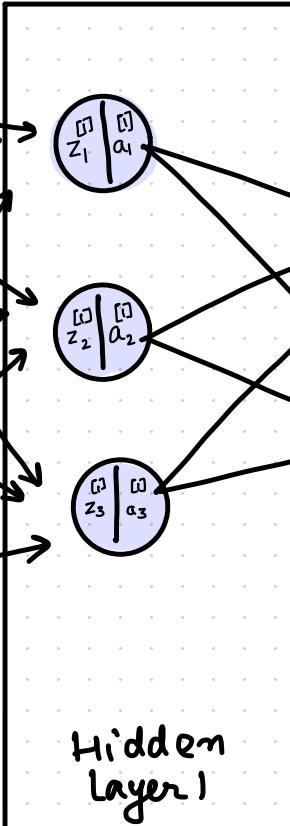




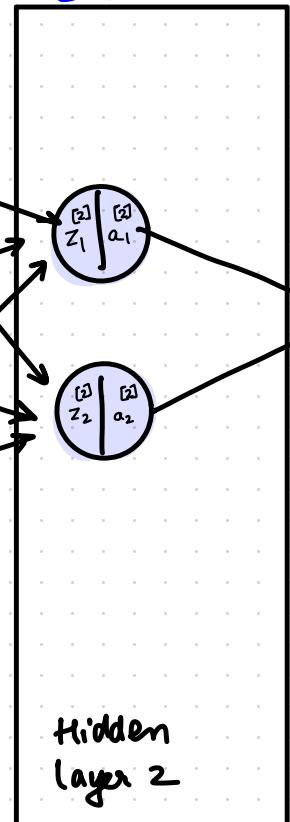
LAYER 0



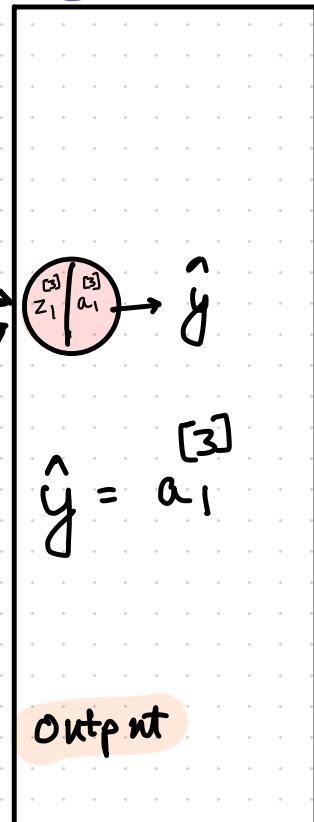
LAYER 1



LAYER 2



LAYER 3

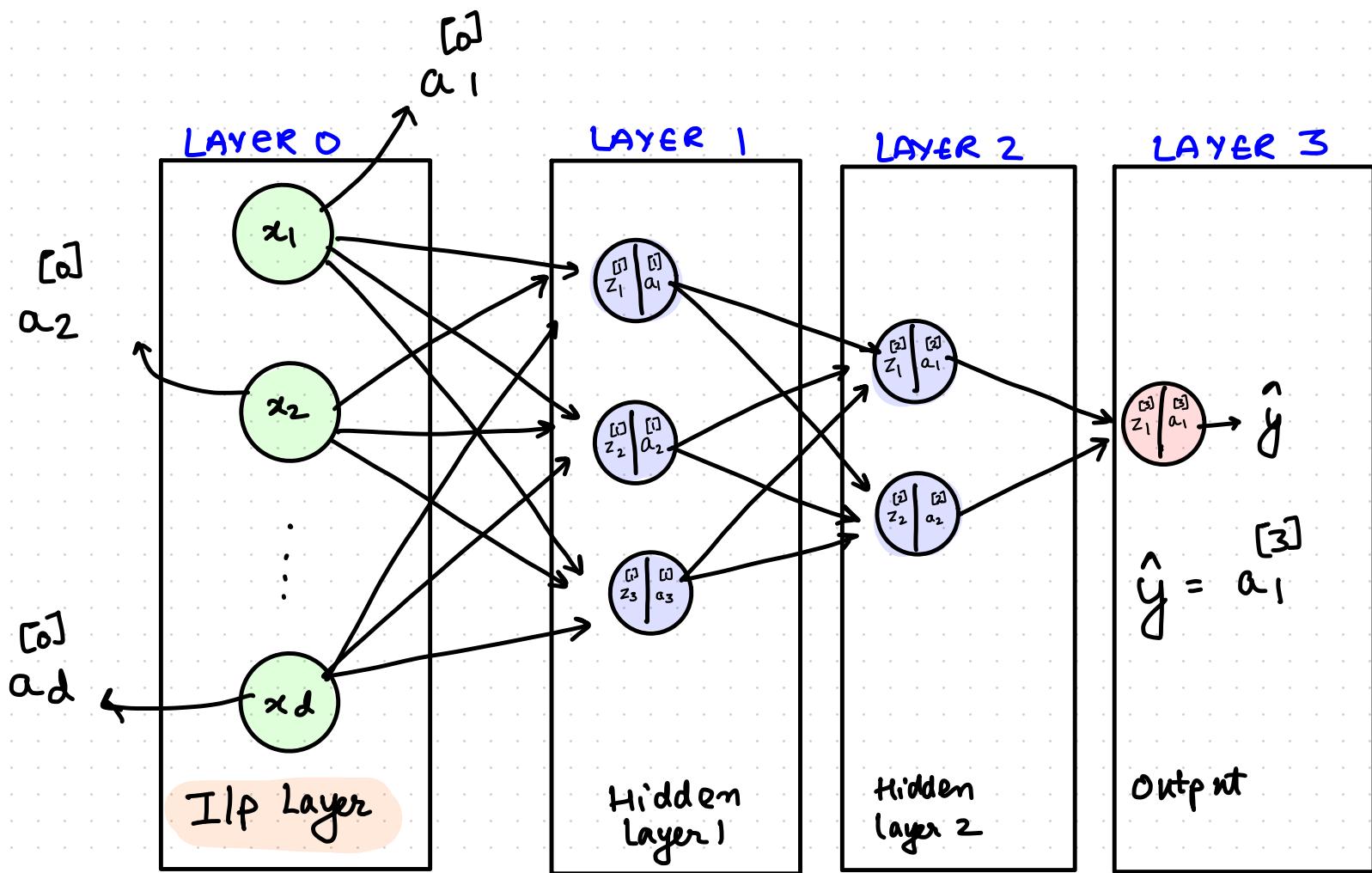


Input Layer

Hidden Layer 1

Hidden Layer 2

Output



IIP =

$$\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

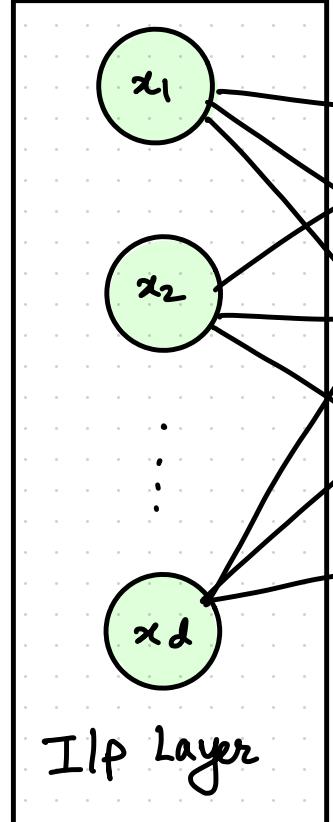
=

$$\begin{bmatrix} [0] \\ a_1 \\ \vdots \\ [0] \\ a_d \end{bmatrix}$$

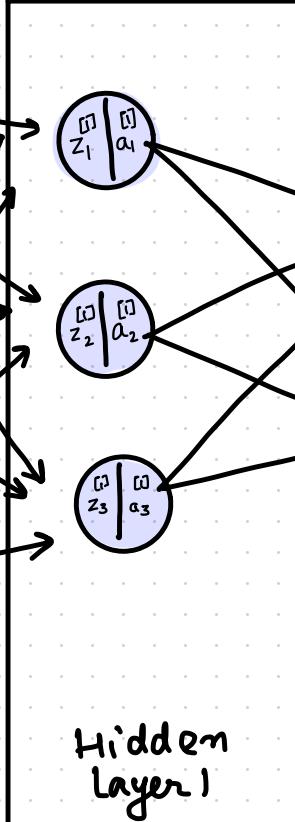
=

$$\begin{bmatrix} [0] \\ a_{d \times 1} \end{bmatrix}$$

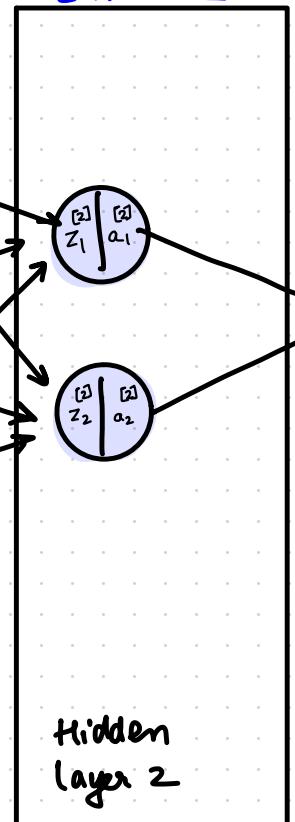
LAYER 0



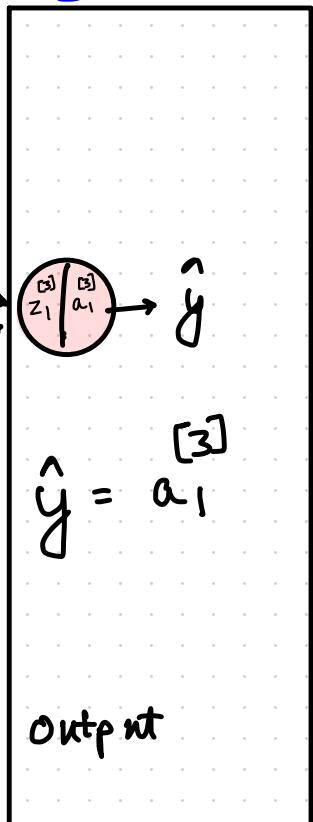
LAYER 1



LAYER 2



LAYER 3



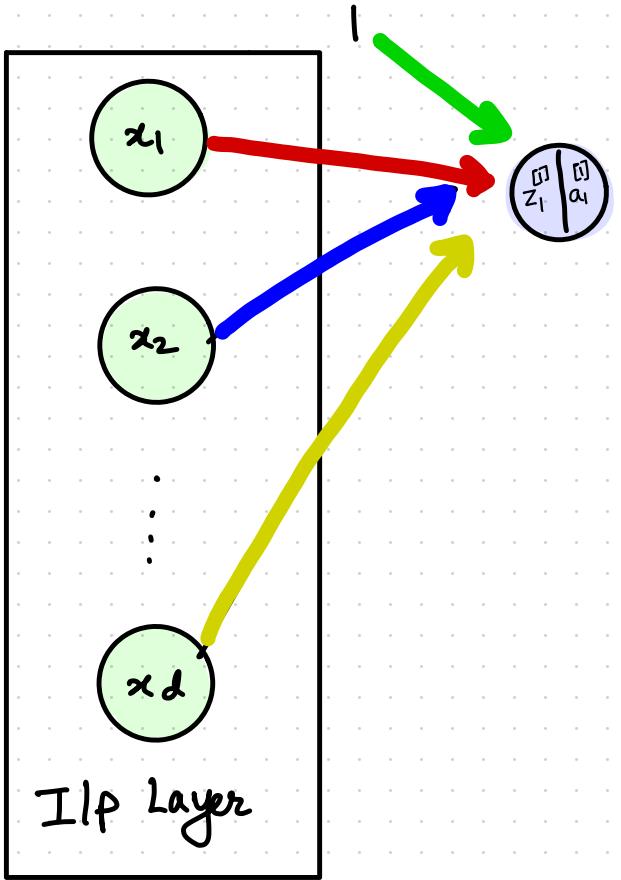
IIP Layer

Hidden Layer 1

Hidden Layer 2

Output

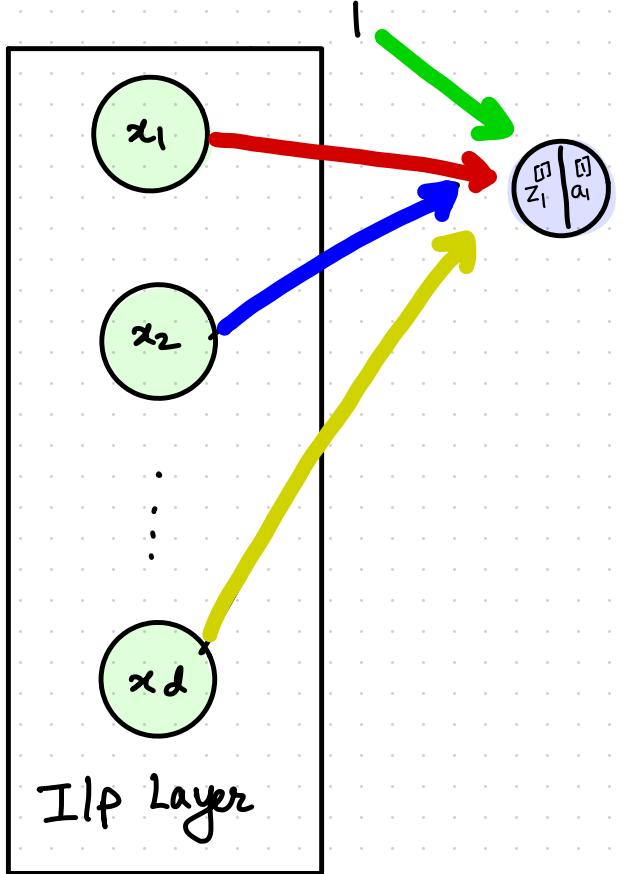
CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



$$z_1^{[1]} = b_1^{[1]} + x_1 * w_{1,1}^{[1]} + x_2 * w_{1,2}^{[1]} + \dots + x_d * w_{1,d}^{[1]}$$

bias layer 1 Node 1

CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



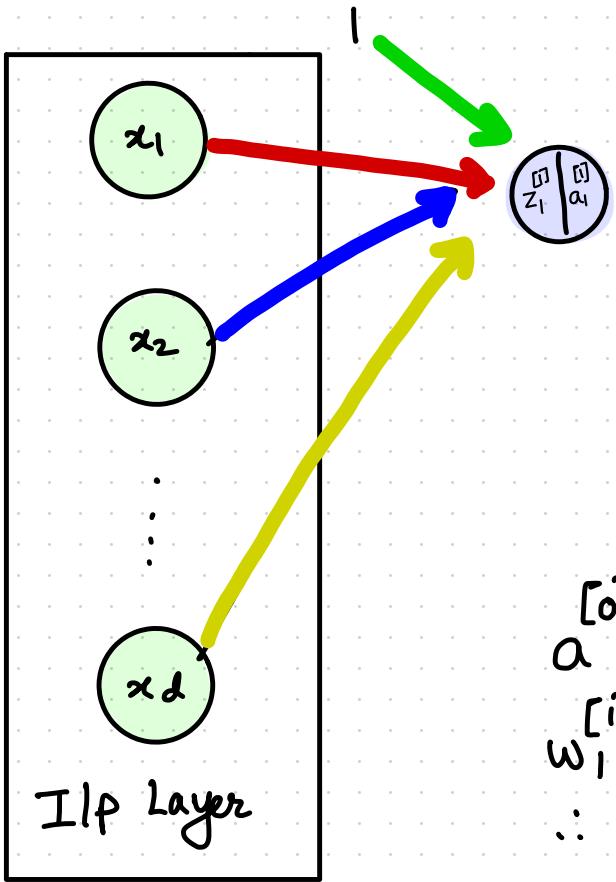
$$z_1^{[1]} = b_1^{[1]} +$$

$$x_1 * w_{1,1}^{[1]} +$$
$$x_2 * w_{1,2}^{[1]} +$$
$$\vdots$$
$$x_d * w_{1,d}^{[1]}$$

bias layer 1 Node 1

$w_{a,b}^{[l]}$ $\leftarrow l^{\text{th}}$ layer
 a^{th} node in l^{th} layer
 b^{th} component of prev. layer activatⁿ

CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



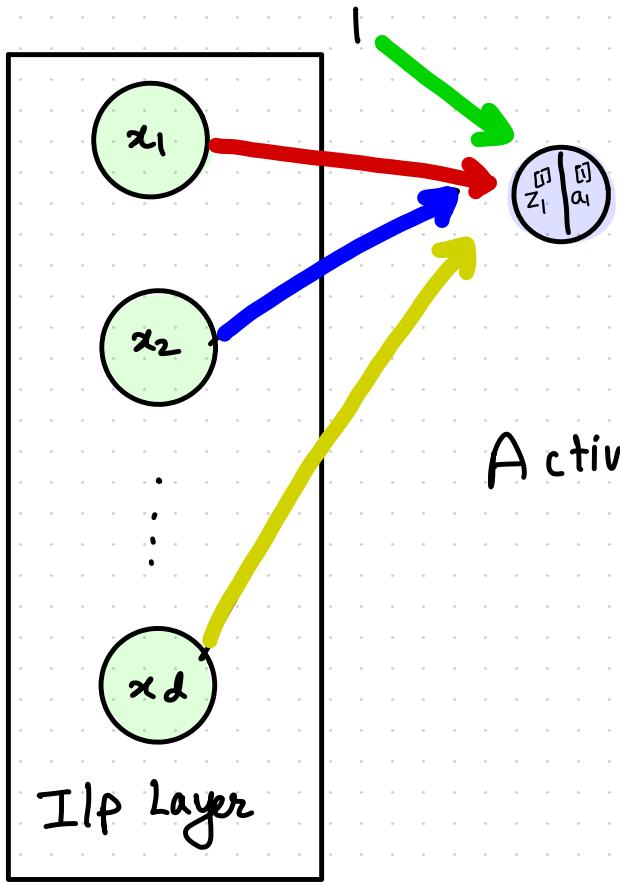
$$z_1^{[i]} = b_1^{[i]} + a_1^{[0]} * w_{1,1}^{[i]} + a_2^{[0]} * w_{1,2}^{[i]} + \dots + a_d^{[0]} * w_{1,d}^{[i]}$$

$$a^{[0]} \in \mathbb{R}^D$$

$$w_1^{[i]} \in \mathbb{R}^D$$

$$\therefore z_1^{[i]} = w_1^{[i]T} a^{[0]} + b_1^{[i]}$$

CONSIDER SINGLE NEURON (LAYER 1, NODE 1)

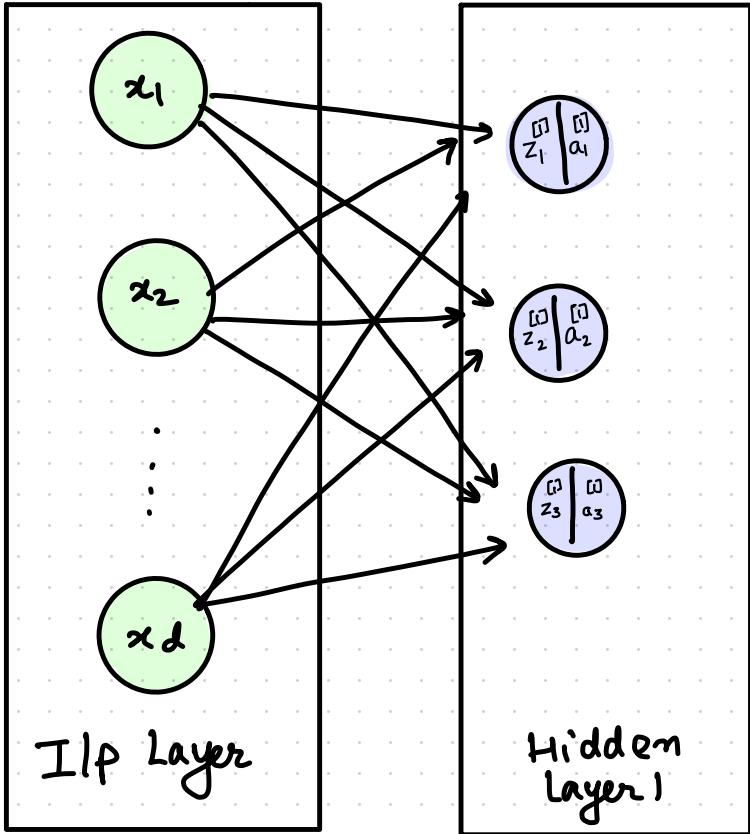


$$z_1^{[1]} = w_1^{[1] T} a_0^{[0]} + b_1^{[1]}$$

$$\text{Activat}^n = a_1^{[1]} = g(z_1^{[1]})$$

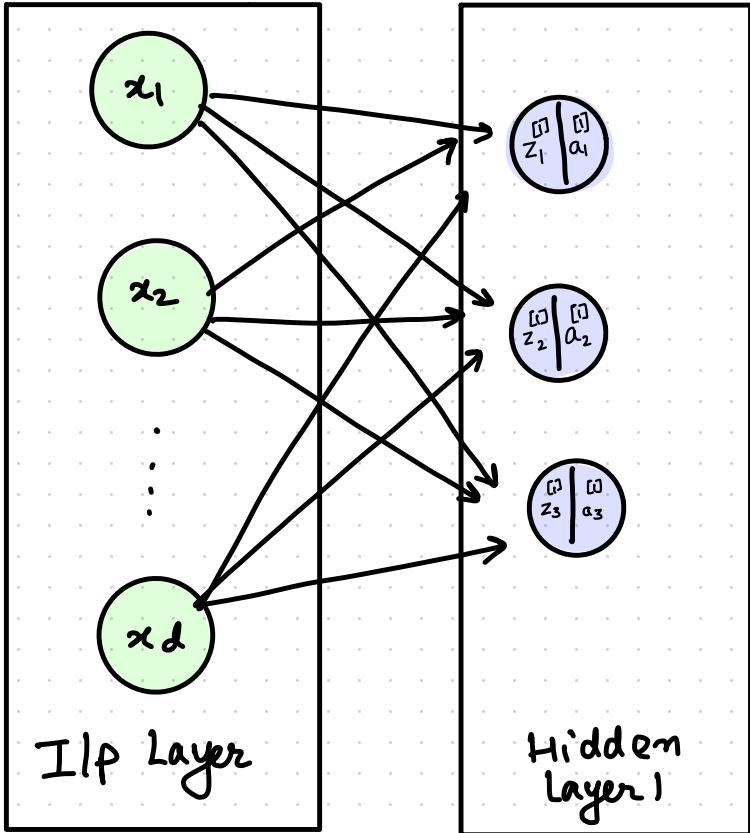
$$a_1^{[1]} \in \mathbb{R}$$

FORWARD PROPAGATION



$$a_i^{[1]} = g\left(\omega_i^{[1] T} a^{[0]} + b_i^{[1]}\right)$$

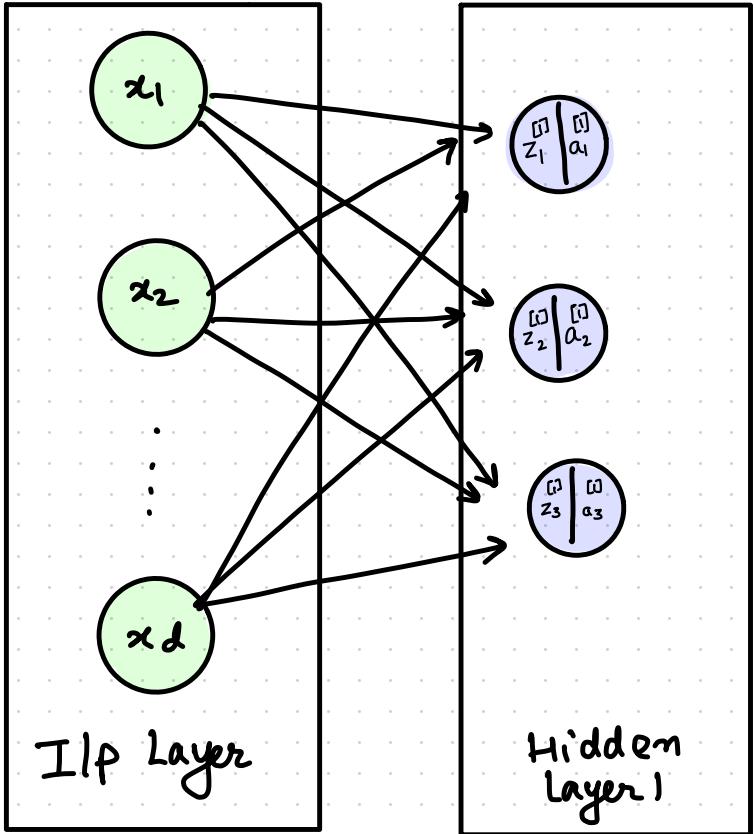
FORWARD PROPAGATION



$$a_1^{[1]} = g\left(\omega_1^{[1]T} a^{[0]} + b_1^{[1]}\right)$$

$$a_2^{[1]} = g\left(\omega_2^{[1]T} a^{[0]} + b_2^{[1]}\right)$$

FORWARD PROPAGATION

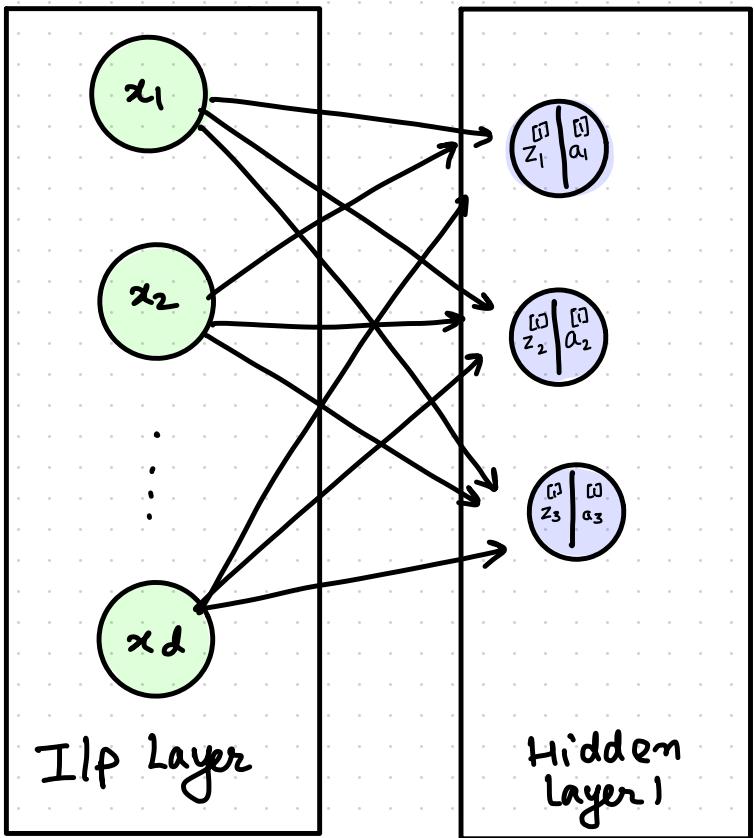


$$a_1^{[1]} = g \left(w_1^{[1] T} a^{[0]} + b_1^{[1]} \right)$$

$$a_2^{[1]} = g \left(w_2^{[1] T} a^{[0]} + b_2^{[1]} \right)$$

$$a_3^{[1]} = g \left(w_3^{[1] T} a^{[0]} + b_3^{[1]} \right)$$

FORWARD PROPAGATION (VECTORISATION)

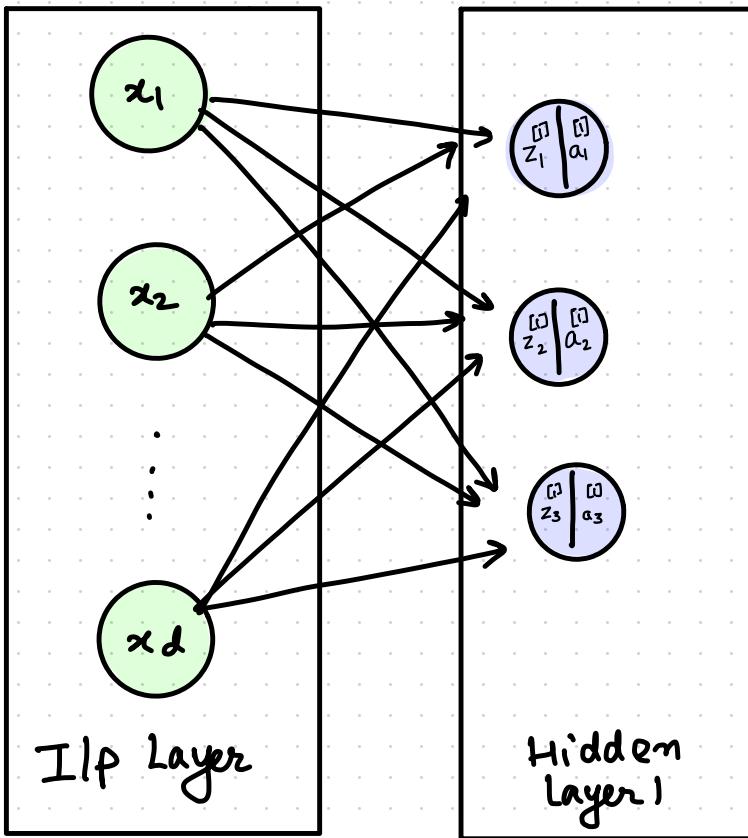


$$z_1^{[1]} = w_1^{[1] T} a^{[0]} + b_1^{[1]}$$

$$z_2^{[1]} = w_2^{[1] T} a^{[0]} + b_2^{[1]}$$

$$z_3^{[1]} = w_3^{[1] T} a^{[0]} + b_3^{[1]}$$

FORWARD PROPAGATION (VECTORISATION)



$$z_1^{[1]} = w_1^{[1] \top} a^{[0]} + b_1^{[1]}$$

Dimensions: $w_1^{[1]}$ is 3×1 , $a^{[0]}$ is 3×1 , $b_1^{[1]}$ is 1×1 .

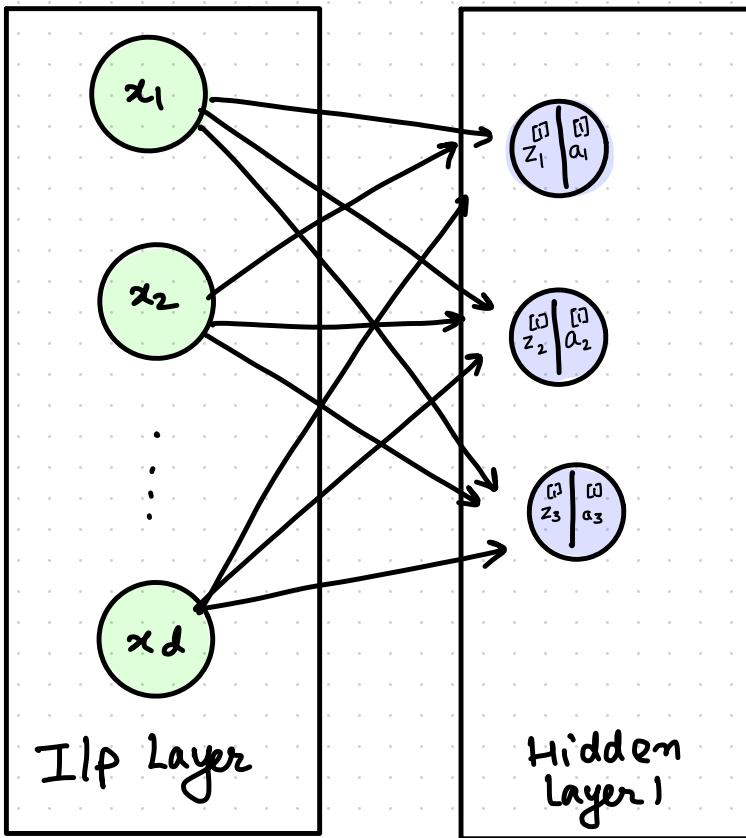
$$z_2^{[1]} = w_2^{[1] \top} a^{[0]} + b_2^{[1]}$$

Dimensions: $w_2^{[1]}$ is 3×1 , $a^{[0]}$ is 3×1 , $b_2^{[1]}$ is 1×1 .

$$z_3^{[1]} = w_3^{[1] \top} a^{[0]} + b_3^{[1]}$$

Dimensions: $w_3^{[1]}$ is 3×1 , $a^{[0]}$ is 3×1 , $b_3^{[1]}$ is 1×1 .

FORWARD PROPAGATION (VECTORISATION)



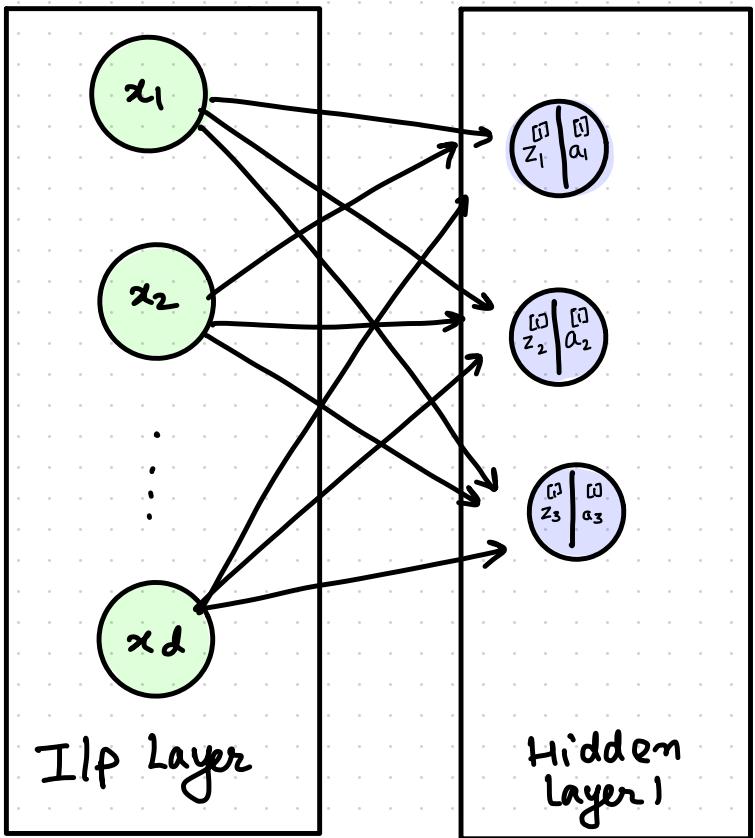
$$z_1^{[1]} = w_1^{[1]T} a^{[0]} + b_1^{[1]} \quad 1 \times 1$$

$$z_2^{[1]} = w_2^{[1]T} a^{[0]} + b_2^{[1]} \quad 1 \times 1$$

$$z_3^{[1]} = w_3^{[1]T} a^{[0]} + b_3^{[1]} \quad 1 \times 1$$

$$z^{[1]} = \begin{bmatrix} -w_1^{[1]T} \\ -w_2^{[1]T} \\ -w_3^{[1]T} \end{bmatrix} a^{[0]} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix} \quad 3 \times 1$$

FORWARD PROPAGATION (VECTORISATION)

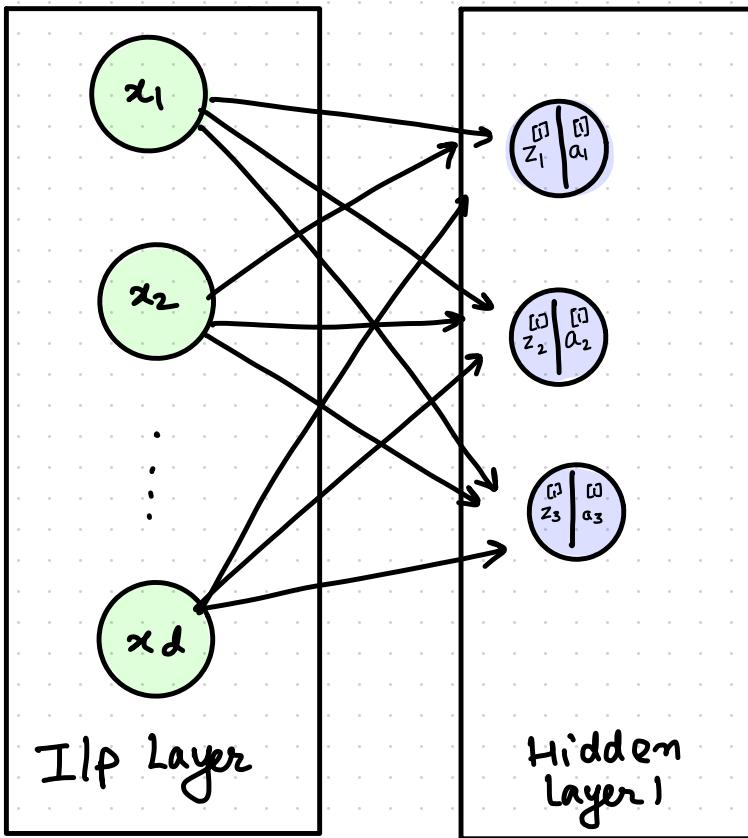


$$z^{[1]}_{3 \times 1} = \begin{bmatrix} -w_1^{[1]T} \\ -w_2^{[1]T} \\ -w_3^{[1]T} \end{bmatrix} a^{[0]}_{3 \times 1} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}_{3 \times 1}$$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]}$$

↑ Capitals for matrices

FORWARD PROPAGATION (VECTORISATION)



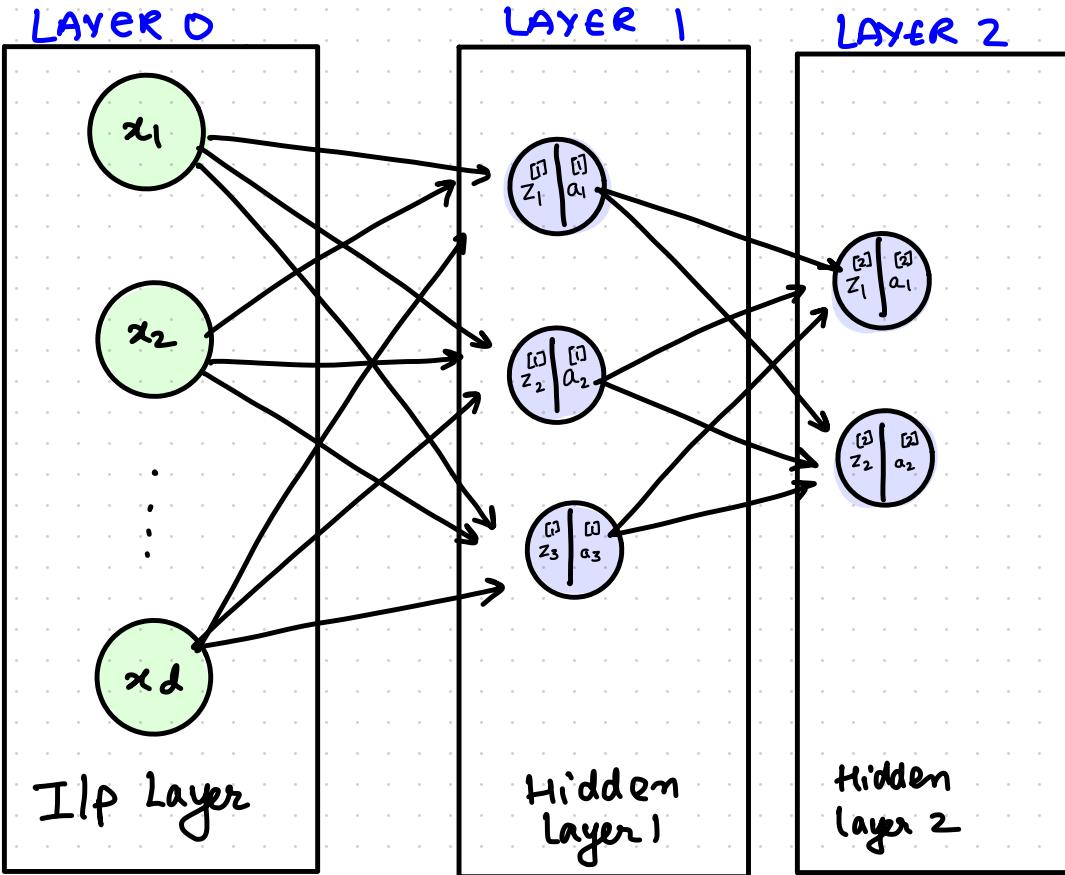
$$z^{[1]}_{3 \times 1} = \begin{bmatrix} -w_1^{[1]T} \\ -w_2^{[1]T} \\ -w_3^{[1]T} \end{bmatrix} a^{[0]}_{3 \times 1} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}_{3 \times 1}$$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]}$$

↑ Capitals for matrices

$$a^{[1]} = g(z^{[1]})$$

FORWARD PROPAGATION

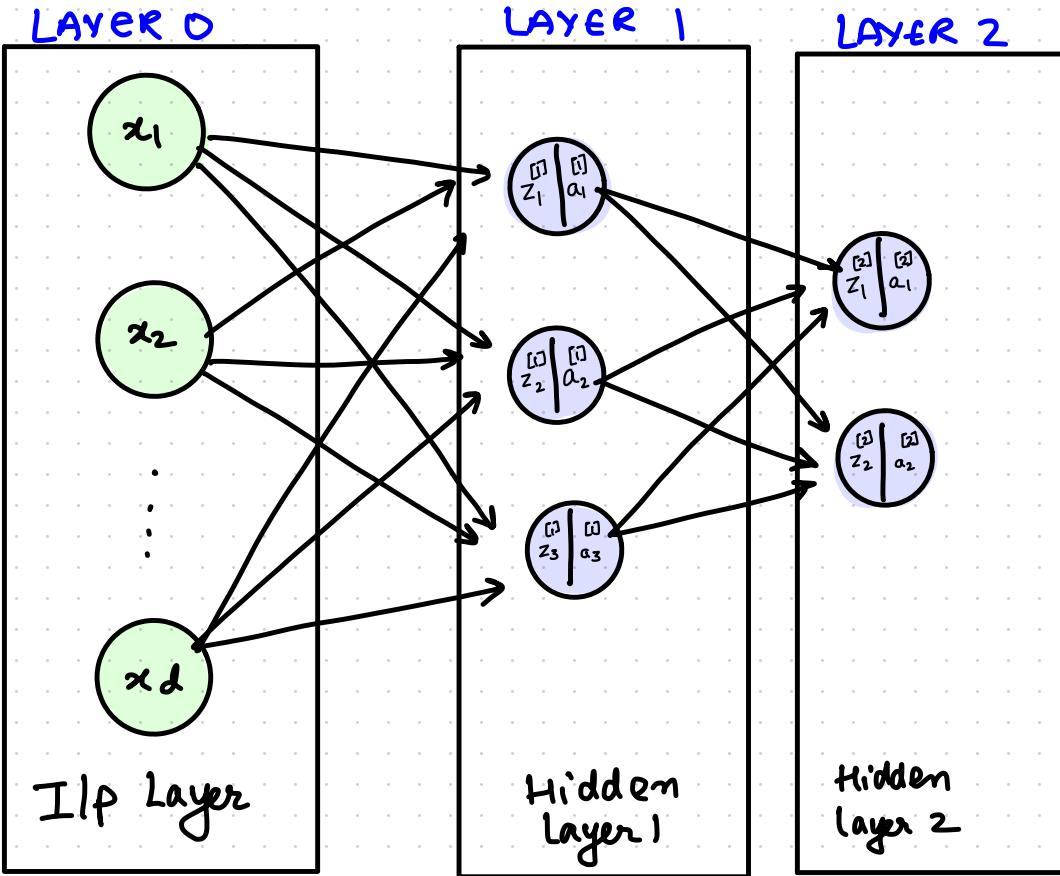


$$z^{[2]} = w^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[2]} = g(z^{[2]})$$

Q. Dim. of $w^{[2]}$?

FORWARD PROPAGATION



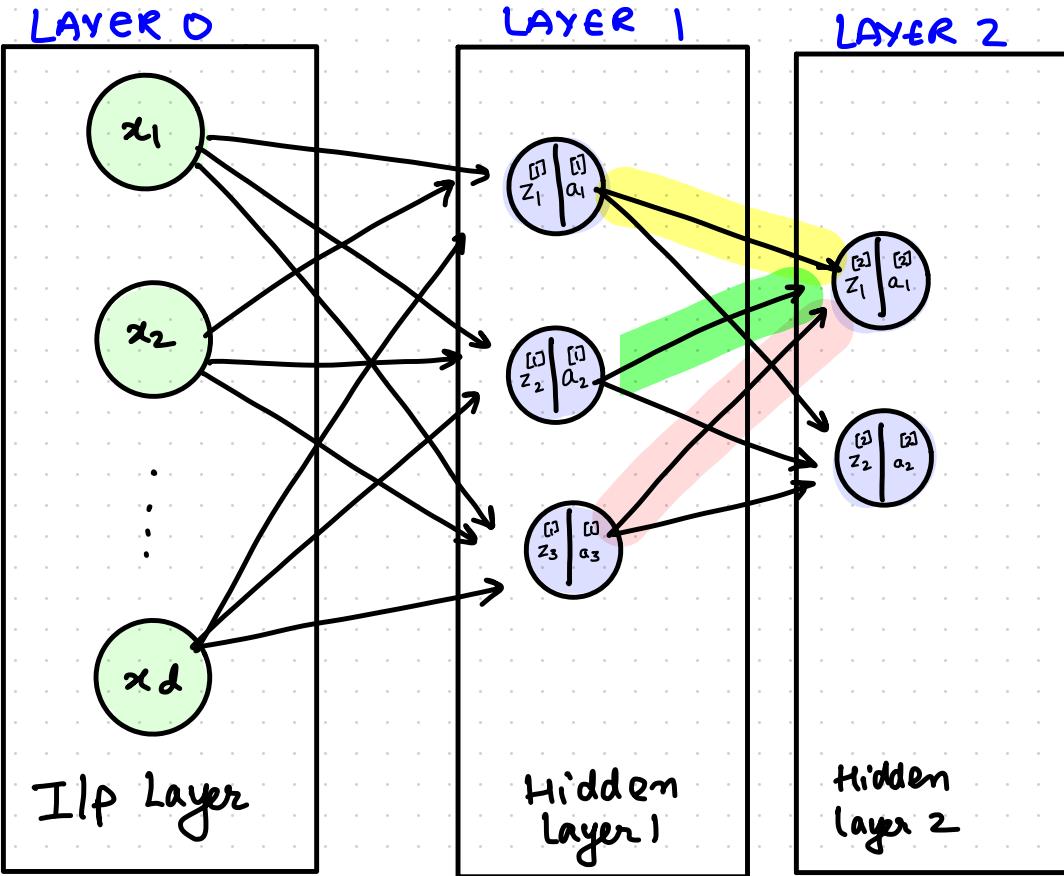
$$z^{[2]} = w^{[2]} a + b^{[2]}$$

$$a^{[2]} = g(z^{[2]})$$

Q. Dim. of $w^{[2]}$?

$$w^{[2]} = \begin{bmatrix} -w_1^{[2]} - \\ -w_2^{[2]} - \end{bmatrix}$$

FORWARD PROPAGATION



$$Z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[2]} = g(Z^{[2]})$$

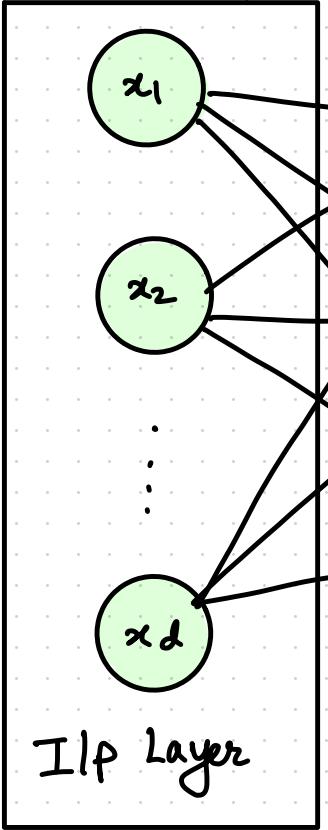
Q. Dim. of $w^{[2]}$?

$$w^{[2]} = \begin{bmatrix} -w_1^{[2]} \\ -w_2^{[2]} \end{bmatrix}^T$$

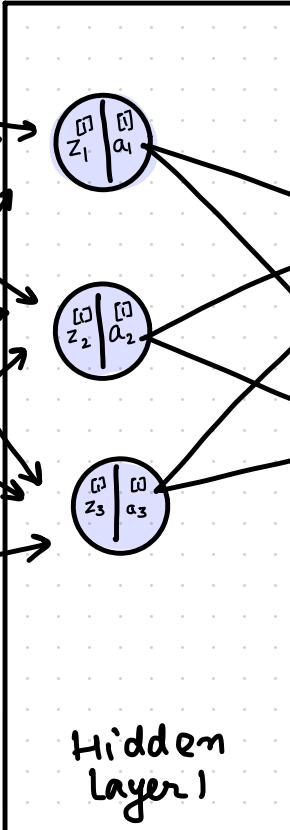
$$\therefore w^{[2]} \in R^{3 \times 2}$$

FORWARD PROPAGATION

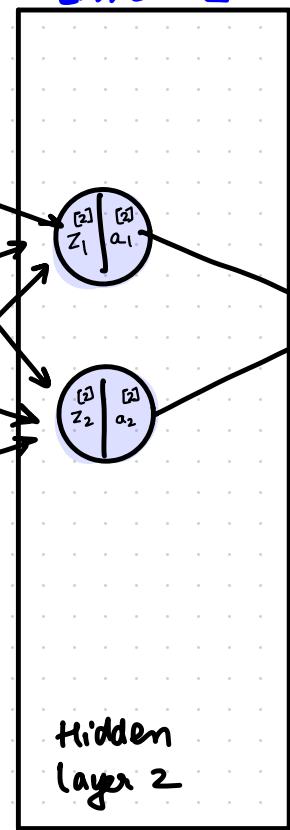
LAYER 0



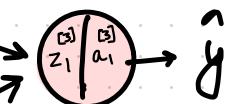
LAYER 1



LAYER 2



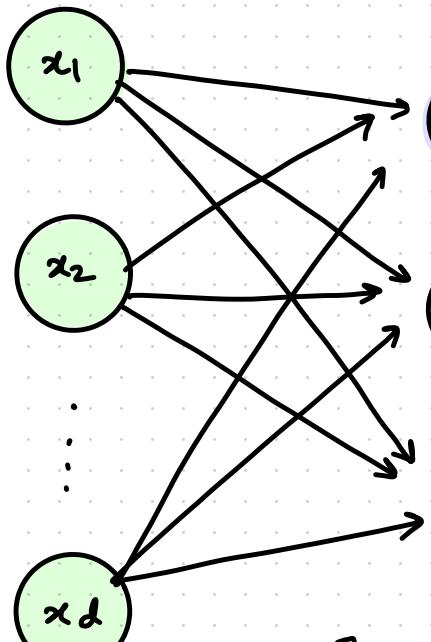
LAYER 3



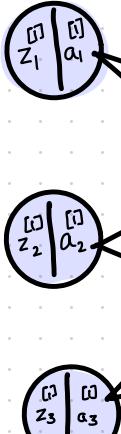
$$\begin{aligned}\hat{y} &= a_1 \\ &= g(z^{[3]}) \\ z^{[3]} &= w^{[3]} a^{[2]} + b^{[3]}\end{aligned}$$

WHAT CAN WE SAY ABOUT SHAPES OF a , b , w

LAYER 0



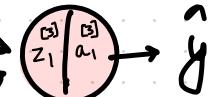
LAYER 1



LAYER 2



LAYER 3



$a^{[0]} \in R^d$ or $R^{N^{[0]}}$
 $d \rightarrow \# \text{ input features}$

$N^{[0]} = \# \text{ units in } 0^{\text{th}} \text{ layer}$

WHAT CAN WE SAY ABOUT SHAPES OF a , b , w

LAYER 0

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d$$

$d \rightarrow \# \text{ input features}$

LAYER 1

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

LAYER 2

LAYER 3

$$\hat{y}$$

$$w^{[1]} = \begin{bmatrix} -\frac{w_1^{[1]}}{}^T \\ -\frac{w_2^{[1]}}{}^T \\ \vdots \\ -\frac{w_{N^{[0]}}^{[1]}}{}^T \end{bmatrix}$$

$$w^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}}$$

WHAT CAN WE SAY ABOUT SHAPES OF a , b , w

LAYER 0

$$x^{[0]} \in \mathbb{R}^d$$

$d \rightarrow \# \text{ input features}$

The diagram shows a layer of input nodes labeled x_1, x_2, \dots, x_d . Arrows connect each x_i to each z_j in the next layer. Below the input layer, it is noted that $a^{[0]} \in \mathbb{R}^d$ and $d \rightarrow \# \text{ input features}$.

LAYER 1

$$z^{[1]} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \in \mathbb{R}^{N^{[1]}}$$

LAYER 2

LAYER 3

$$\hat{y} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \end{bmatrix}$$

SUMMARY OF SHAPES

$N^{[l]}$: # nodes in l^{th} layer

$$a^{[0]} \in \mathbb{R}^{N^{[0]}}$$

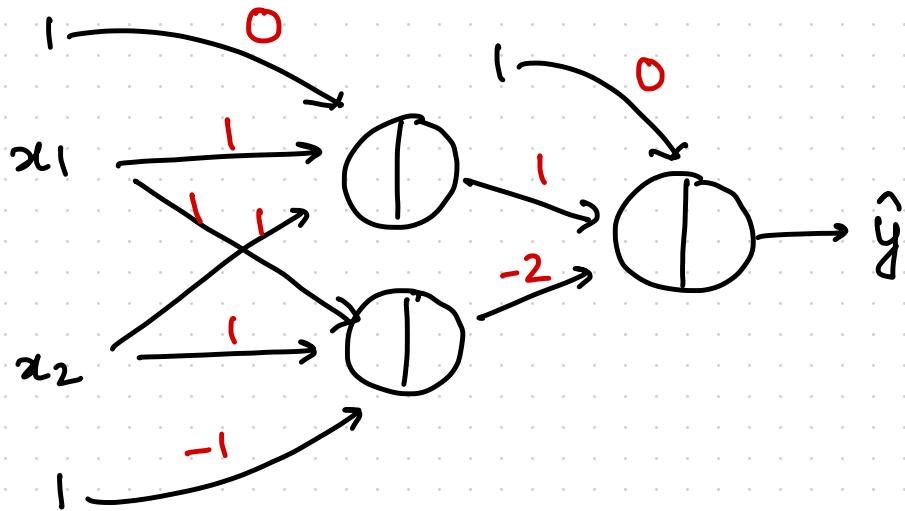
$$w^{[l]} \in \mathbb{R}^{N^{[l]} \times N^{[l-1]}}$$

$$b^{[l]} \in \mathbb{R}^{N^{[l]}}$$

$$z^{[l]} \in \mathbb{R}^{N^{[l]}}$$

$$\alpha^{[l]} \in \mathbb{R}^{N^{[l]}}$$

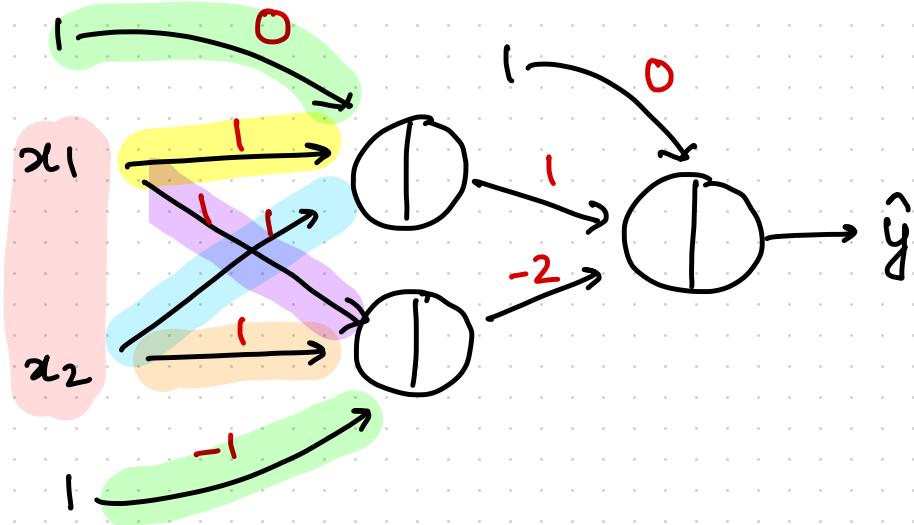
XOR USING "MLP" RELU



CONFIRM If ABOVE N/W IS CORRECT FOR XOR.

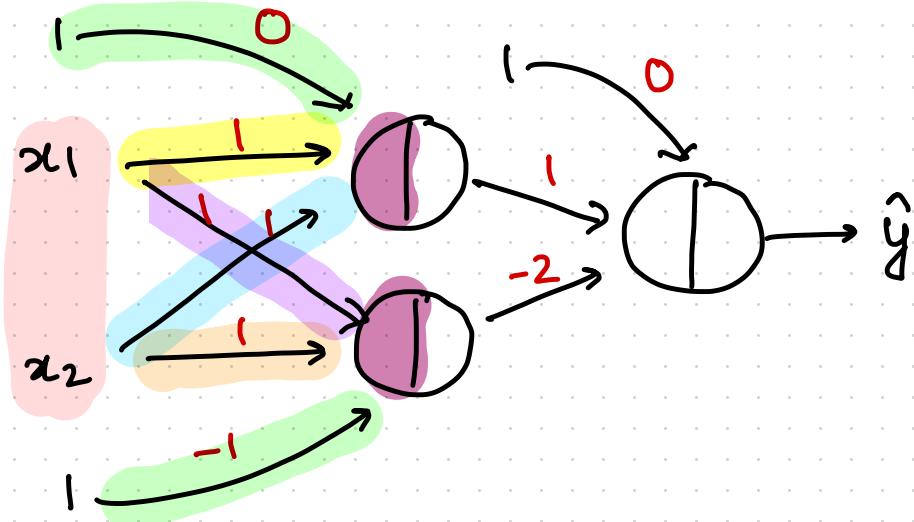
Start with $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $y_{\text{true}} = 0$

XOR USING "MLP" RELU



$$a^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; b^{[0]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}; W^{[1]} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

XDR USING "MLP" RELU



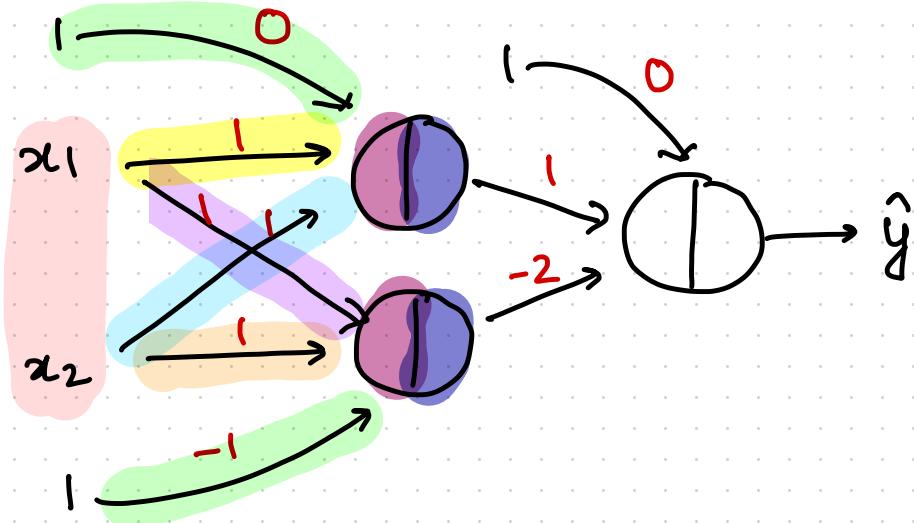
$$a^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix};$$

$$b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix};$$

$$W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

XOR USING "MLP" RELU



$$a^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

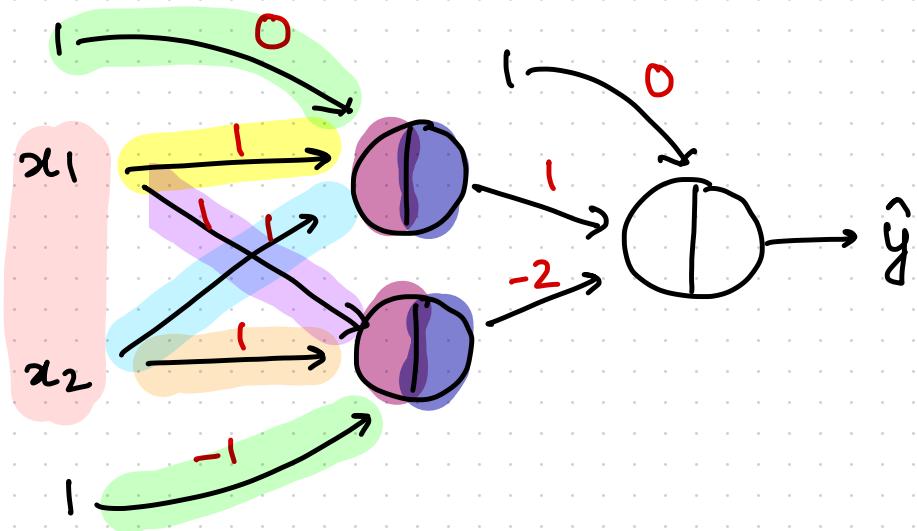
$$b^{[0]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$a^{[1]} = \text{RELU} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

XOR USING "MLP" RELU

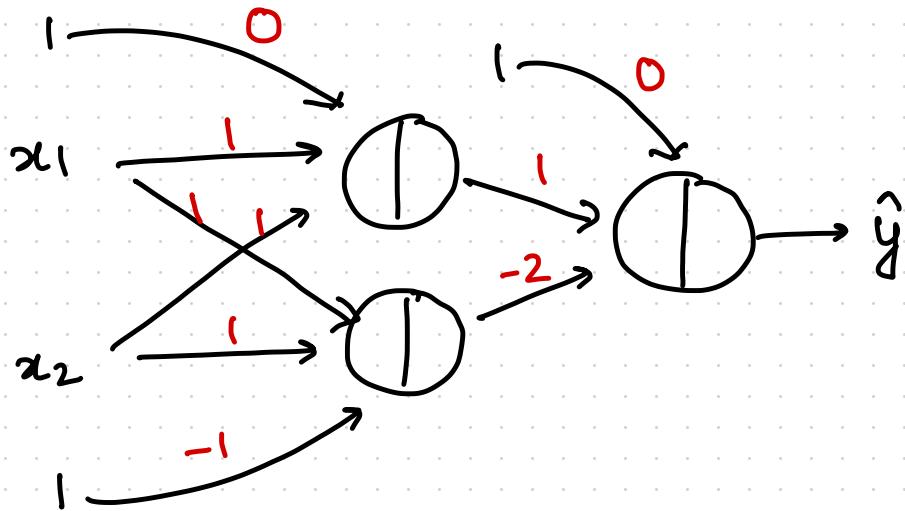


$$a^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} ; W^{[2]} = \begin{bmatrix} 1 & -2 \end{bmatrix} ; b^{[2]} = \begin{bmatrix} 0 \end{bmatrix}$$

$$z^{[2]} = [1 \ -2] \begin{bmatrix} 0 \\ 0 \end{bmatrix} + [0] = 0 ; a^{[2]} = \hat{y} = \text{RELU}(0) = 0$$

✓

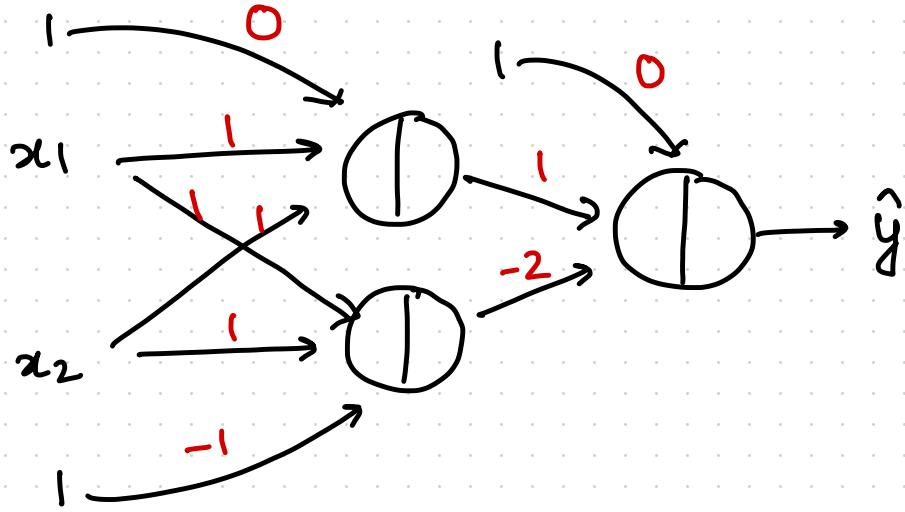
XOR USING "MLP" RELU



CONFIRM If ABOVE n/w IS CORRECT FOR XOR.

Start with $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $y_{\text{true}} = 1$

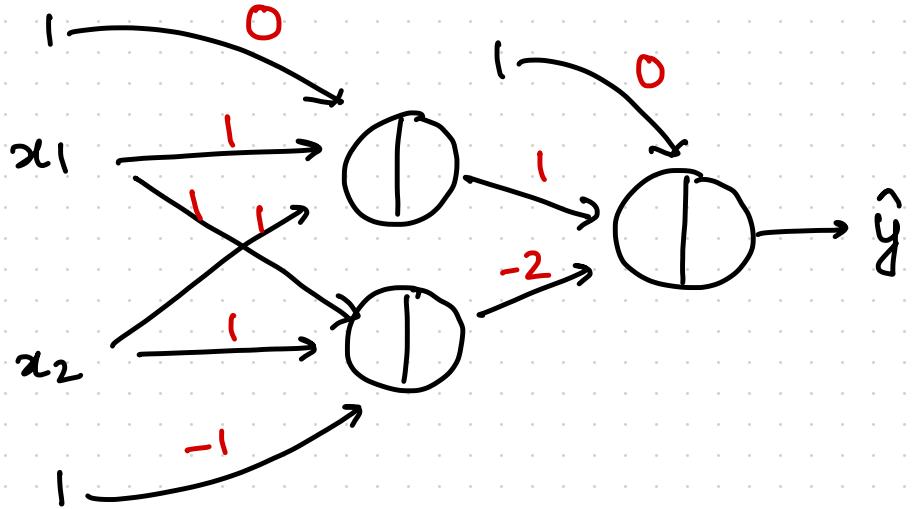
XDR USING "MLP" RELU



$$z^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$a^{[1]} = \text{RELU}(z^{[1]}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

XOR USING "MLP" RELU



$$z^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$a^{[1]} = \text{RELU}(z^{[1]}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$z^{[2]} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix}$$

$$\begin{aligned} a^{[2]} &= \text{RELU}(1) = 1 \\ &= \hat{y} \end{aligned}$$

COMPUTATION FOR N EXAMPLES

$x_{(i)} \in R^d$ or $R^{N^{[0]}}$

$$X = \begin{bmatrix} -x_0^T - \\ -x_1^T - \\ -x_2^T - \\ -x_N^T - \end{bmatrix} = \begin{bmatrix} -a_{(0)}^{[0]} - \\ \vdots \\ -a_{(n)}^{[0]} - \end{bmatrix} = A^{[0]} \in R^{N \times N^{[0]}}$$

↑ matrix

COMPUTATION FOR N EXAMPLES

$x_{(i)} \in R^d$ or $R^{N^{[0]}}$

$$X = \begin{bmatrix} -x_0^T - \\ -x_{(1)}^T - \\ -x_{(2)}^T - \\ -x_{(N)}^T - \end{bmatrix} = \begin{bmatrix} -a_{(0)}^{[0]}^T - \\ \vdots \\ -a_{(N)}^{[0]}^T - \end{bmatrix} = A^{[0]} \in R^{N \times N^{[0]}}$$

$Z^{[i]} \leftarrow$ Layer
 $z_{(i)} \leftarrow$ Instance/
 Sample

$$= W^{[i]} A^{[0]}_{(i)}^T + b^{[i]}$$

$\in R^{N^{[i]}}$

Independent of "i"

$a_{(i)}^{[0]} \in R^d \equiv [::]$

COMPUTATION FOR N EXAMPLES

$$x_{(i)} \in R^D$$

$$X = \begin{bmatrix} -x_0^T - \\ -x_{(1)}^T - \\ -x_{(2)}^T - \\ \vdots \\ -x_{(N)}^T - \end{bmatrix} = \begin{bmatrix} -a_{(0)}^{[0]} - \\ \vdots \\ -a_{(N)}^{[0]} - \end{bmatrix}^T$$

$$z_{(1)}^{[1]} = \left\{ W a_{(1)}^{[0]} + b^{[1]} \right\} \in R^{N^{[1]}}$$

$$z_{(2)}^{[1]} = \left\{ W a_{(2)}^{[0]} + b^{[1]} \right\} \in R^{N^{[1]}}$$

$$\vdots$$

$$z_{(N)}^{[1]} = \left\{ W a_{(N)}^{[0]} + b^{[1]} \right\} \in R^{N^{[1]}}$$

COMPUTATION FOR N EXAMPLES

$$x_{(i)} \in R^D$$

$$X = \begin{bmatrix} -x_{(1)}^T - \\ -x_{(2)}^T - \\ -x_{(N)}^T - \end{bmatrix} = \begin{bmatrix} -a_{(1)}^{[0]}^T - \\ \vdots \\ -a_{(N)}^{[0]}^T - \end{bmatrix} = A \in R^{N \times N^{[0]}}$$

$$z_{(1)}^{[1]} = w^{[1]} a_{(1)}^{[0]} + b^{[1]}$$

$\in R^{N^{[1]}}$

$$z_{(2)}^{[1]} = w^{[1]} a_{(2)}^{[0]} + b^{[1]}$$

$\in R^{N^{[1]}}$

$$\vdots$$

$$z_{(N)}^{[1]} = w^{[1]} a_{(N)}^{[0]} + b^{[1]}$$

$\in R^{N^{[1]}}$

COMPUTATION FOR N EXAMPLES

$$x_{(i)} \in R^D$$

$$X = \begin{bmatrix} -x_{(1)}^T - \\ -x_{(2)}^T - \\ \vdots \\ -x_{(N)}^T - \end{bmatrix} = \begin{bmatrix} -a_{(1)}^{[0]}^T - \\ \vdots \\ -a_{(N)}^{[0]}^T - \end{bmatrix} = A \in R^{N \times N^{[0]}}$$

$$z_{(1)}^{[1]} = w^{[1]} a_{(1)}^{[0]} + b^{[1]}$$

$$z_{(2)}^{[1]} = w^{[1]} a_{(2)}^{[0]} + b^{[1]}$$

$$\vdots$$

$$z_{(N)}^{[1]} = w^{[1]} a_{(N)}^{[0]} + b^{[1]}$$

$$\Rightarrow Z = \begin{bmatrix} -z_{(1)}^{[1]} - \\ -z_{(2)}^{[1]} - \\ \vdots \\ -z_{(N)}^{[1]} - \end{bmatrix} \in R^{N \times N^{[1]}}$$

COMPUTATION FOR N EXAMPLES

$$x_{(i)} \in R^D$$

$$X = \begin{bmatrix} -x_{(1)}^T - \\ -x_{(2)}^T - \\ \vdots \\ -x_{(N)}^T - \end{bmatrix} = \begin{bmatrix} -a_{(1)}^{[0]}^T - \\ \vdots \\ -a_{(N)}^{[0]}^T - \end{bmatrix} = A \in R^{N \times N^{[0]}}$$

$$z_{(1)}^{[1]} = w^{[1]} a_{(1)}^{[0]} + b^{[1]}$$

$$z_{(2)}^{[1]} = w^{[1]} a_{(2)}^{[0]} + b^{[1]}$$

$$\vdots$$

$$z_{(N)}^{[1]} = w^{[1]} a_{(N)}^{[0]} + b^{[1]}$$

$$\Rightarrow Z^{[1]} = \begin{bmatrix} -z_{(1)}^{[1]} - \\ -z_{(2)}^{[1]} - \\ \vdots \\ -z_{(N)}^{[1]} - \end{bmatrix} \in R^{N \times N^{[1]}}$$

COMPUTATION FOR N EXAMPLES

$$A^{[0]} \in \mathbb{R}^{N \times N^{[0]}}$$

$$W^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}}$$

$$b^{[1]} \in \mathbb{R}^{N^{[1]}}$$

$$B^{[1]} = \begin{bmatrix} -b^{[1]T} \\ \vdots \\ -b^{[1]T} \end{bmatrix} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} \in \mathbb{R}^{N \times N^{[1]}}$$

all same entries

COMPUTATION FOR N EXAMPLES

$$A^{[0]} \in \mathbb{R}^{N \times N^{[0]}}$$

$$W^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}}$$

$$b^{[1]} \in \mathbb{R}^{N^{[1]}}$$

$$B^{[1]} = \begin{bmatrix} -b^{[1]T} \\ \vdots \\ -\bar{b}^{[1]T} \end{bmatrix} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} = A^{[0]} W^{[1]} + B^{[1]}$$

COMPUTATION FOR N EXAMPLES

$$A^{[0]} \in R^{N \times N^{[0]}}$$

$$W^{[1]} \in R^{N^{[1]} \times N^{[0]}}$$

$$b^{[1]} \in R^{N^{[1]}}$$

$$B^{[1]} = \begin{bmatrix} -b^{[1]T} \\ \vdots \\ -\bar{b}^{[1]T} \end{bmatrix} \in R^{N \times N^{[1]}}$$

$$Z^{[1]} \in R^{N \times N^{[1]}}$$

$$Z^{[1]} = A^{[0]} W^{[1]T} + B^{[1]} \Rightarrow A^{[1]} = g(Z^{[1]})$$

$$\boxed{Z^{[l]} = A^{[l-1]} W^{[l]} + B^{[l]}}$$

$$A^{[l]} = g(Z^{[l]})$$

XOR ALL EXAMPLES

① $x = A^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}_{4 \times 2}$; $\hat{y} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \end{bmatrix}_{4 \times 1}$

XOR ALL EXAMPLES

$$\textcircled{1} \quad X = A^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}_{4 \times 2} ; \quad \hat{y} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \end{bmatrix}_{4 \times 1}$$

$$w^{[1]} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}_{2 \times 2} ; \quad b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \Rightarrow B^{[1]} = \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}_{4 \times 2}$$

XOR ALL EXAMPLES

$$\textcircled{1} \quad X = A^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}_{4 \times 2} ; \quad \hat{y} = \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix}_{3 \times 1}$$

$$W^{[1]} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}_{2 \times 2} = W^{[1]T} ; \quad b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \Rightarrow B^{[1]} = \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}_{4 \times 2}$$

$$Z^{[1]} = A^{[0]} W^{[1]T} + B^{[1]} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \\ 2 & 1 \end{bmatrix}$$

XOR ALL EXAMPLES

$$Z^{[1]} = A^{[0]} W^{[1]} + B^{[1]} = \begin{bmatrix} 0 & 0 \\ -1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & -1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \\ 0 & 0 \\ 2 & -1 \end{bmatrix}$$

$$A^{[1]} = \text{RELU}\left(\begin{bmatrix} 0 & -1 \\ 1 & 0 \\ -1 & 0 \\ 2 & 1 \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}_{4 \times 2}$$

$$W^{[2]} = \begin{bmatrix} 1 & -2 \end{bmatrix}$$

$$b^{[2]} = [0] \Rightarrow B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

XOR ALL EXAMPLES

$$A^{[1]} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}_{4 \times 2}$$

$$w^{[2]} = [1 \ -2]_{1 \times 2}$$

$$b^{[2]} = [0] \Rightarrow B^{[2]} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{4 \times 1}$$

$$z^{[2]} = A^{[1]} w^{[2]} + B^{[2]} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

XOR ALL EXAMPLES

$$z^{[2]} = A^{[1]} \omega^{[2]^\top} + b^{[2]} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ -1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$

$$A^{[2]} = \hat{y} = \text{RELU} \left(\begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} \right) = y_{G.T}$$

PARAMETERS

Parameters: $w^{[l]}$; $b^{[l]}$ $\neq l$

Q: # parameters for XOR example?

PARAMETERS

Parameters: $w^{[l]}$; $b^{[l]}$ $\neq l$

Q: # parameters for XOR example?

$$N^{[0]} = 2; N^{[1]} = 2; N^{[2]} = 1$$

$$w^{[1]} \in R^{N^{[0]} \times N^{[1]}} = \begin{bmatrix} & \\ & \end{bmatrix}_{2 \times 2} = 4 \text{ params} = N^{[1]} * N^{[0]}$$

$$b^{[1]} \in R^{N^{[1]}} = \begin{bmatrix} & \\ & \end{bmatrix}_{2 \times 1} = 2 \text{ params} = N^{[1]}$$

$$w^{[2]} \rightarrow 2 \text{ params} = N^{[2]} * N^{[1]} \quad \& \quad b^{[2]} = N^{[2]} \text{ params}$$

PARAMETERS

Parameters: $w^{[l]}$; $b^{[l]}$ $\forall l$

$$\text{PARAMS} = \sum_{l=1}^L N^{[l]} \cdot N^{[l-1]} + N^{[l]}$$

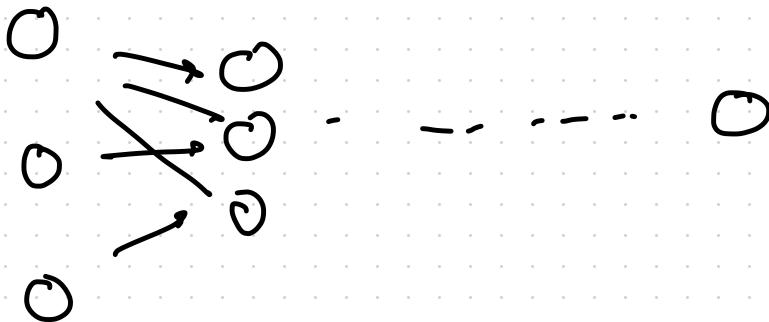
? XOR 4 examples, 9 parameters ??

⇒ NOTEBOOK: XOR demo

LEARNING

PARAMETERS

FORWARD PROPAGATION (PREDICT BASED ON CURRENT PARAMS)



BACKWARD PROPAGATION (CHANGE WEIGHTS TO IMPROVE OBJECTIVE)

LEARNING

PARAMETERS

Assume $\text{if } p : X \in \mathbb{R}^{N \times N^{[0]}}$

olp : \hat{y} ; G.T. : y

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N L(\hat{y}^{(i)}, y)$$

$$\text{or } \sum_{i=1}^N L(\hat{y}^{(i)}, y)$$

Params: $w^{[i]}, b^{[i]}, \dots$

LEARNING PARAMETERS

Assume $\text{if } p : X \in \mathbb{R}^{N \times N^{[C]}}$

$\text{olp} : \hat{y}; \text{ G.T.} : y$

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N L(\hat{y}^{(i)}, y) = J(\theta)$$

Params: $\theta = \{w^{[1]}, b^{[1]}, \dots\}$

GRADIENT DESCENT

① INIT Params randomly

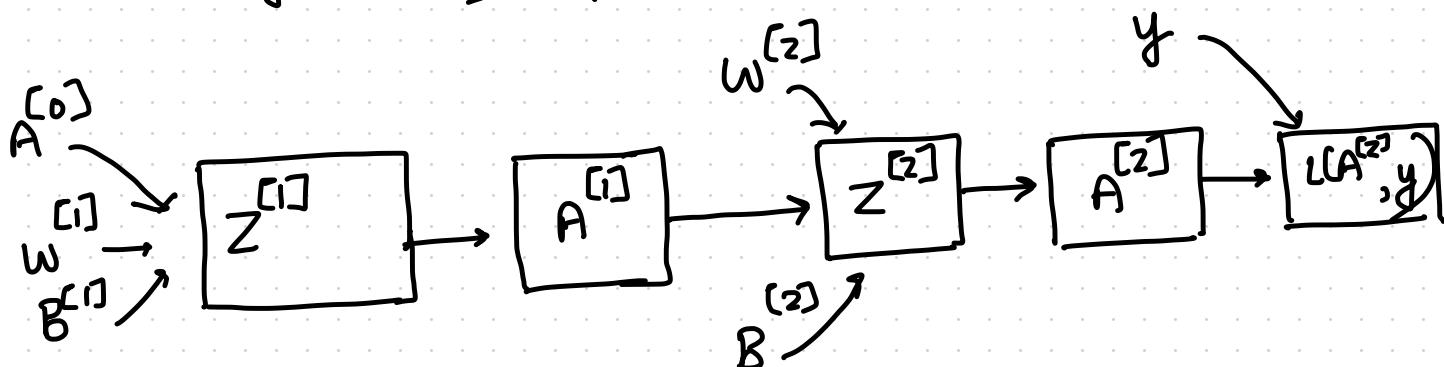
② Till convergence:

$$w^{[1]} = w^{[1]} - \alpha \frac{\partial J(\theta)}{\partial w^{[1]}}$$

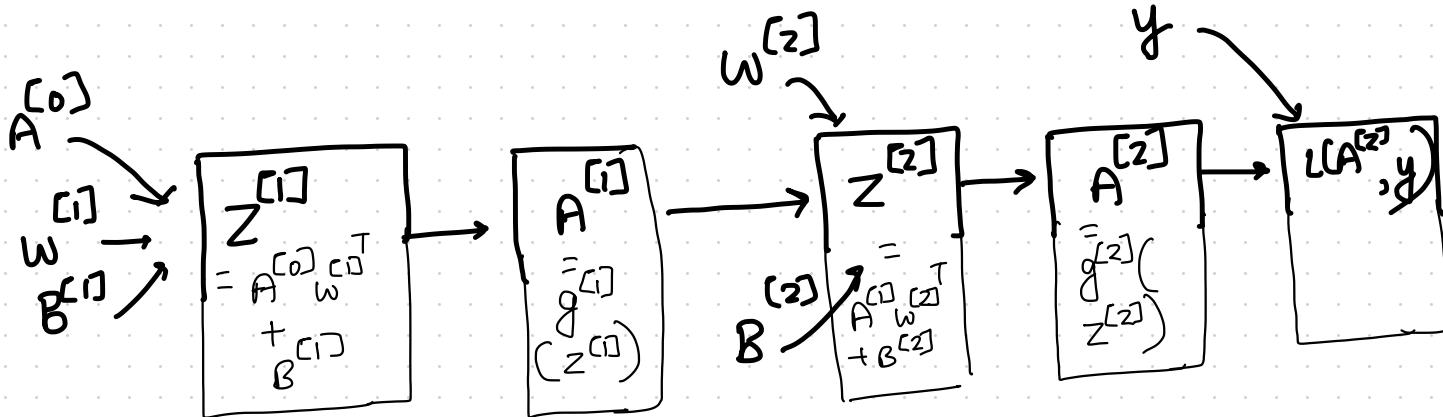
How TO COMPUTE?

COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$\begin{aligned} \textcircled{1} \quad z^{[1]} &= A^{[0]} w^{[1]} + b^{[1]} \\ \textcircled{2} \quad A^{[1]} &= g^{[1]}(z^{[1]}) \\ \textcircled{3} \quad z^{[2]} &= A^{[1]} w^{[2]} + b^{[2]} \\ \textcircled{4} \quad A^{[2]} &= g^{[2]}(z^{[2]}) = \hat{y} \end{aligned}$$



COMPUTATION GRAPH (FOR XOR EXAMPLE)



Let $g^{[1]} = \text{SIGMOID}$; ASSUME CROSS ENTROPY LOSS
 $g^{[2]} = \text{SIGMOID}$

WHAT IS $\frac{\partial L(\theta)}{\partial w^{[0]}}$? ; WHAT IS $\frac{\partial L(\theta)}{\partial A^{[2]}}$?

DERIVATIVES OF ACTIVATION FUNCTIONS

RELU

$$g(z) = \begin{cases} z; & z > 0 \\ 0; & z < 0 \\ \text{undefined}; & 0/w \end{cases}$$

↓ Assume $z \neq 0$

$$g(z) = \begin{cases} z; & z > 0 \\ 0; & z \leq 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z > 0 \\ 0; & 0/w \end{cases}$$

DERIVATIVES OF ACTIVATION FUNCTIONS

RELU (LEAKY)

$$g(z) = \begin{cases} z; & z > 0 \\ \alpha z; & z \leq 0 \\ \text{undefined}; & 0/w \end{cases}$$



$$g(z) = \begin{cases} z; & z > 0 \\ \alpha z; & z \leq 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z > 0 \\ \alpha; & z \leq 0 \end{cases}$$

DERIVATIVES OF ACTIVATION FUNCTIONS

SIGMOID

$$g(z) = \frac{1}{1+e^{-z}}$$

$$g'(z) = \frac{-1}{(1+e^{-z})^2} \frac{d}{dz} (1+e^{-z}) = \frac{-1 (e^{-z})(-1)}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$g'(z) = \frac{1+e^{-z}}{(1+e^{-z})^2} - \frac{1}{(1+e^{-z})^2} = g(z)(1-g(z))$$

DERIVATIVES OF ACTIVATION FUNCTIONS

TANH

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{u}{v}$$

$$g'(z) = \frac{v \frac{du}{dz} - u \frac{dv}{dz}}{v^2} = \frac{\left(\frac{z}{e} + \frac{-z}{e^{-z}}\right)\left(\frac{z}{e} + \frac{-z}{e^{-z}}\right) - \left(\frac{z}{e} - \frac{-z}{e^{-z}}\right)\left(\frac{z}{e} - \frac{-z}{e^{-z}}\right)}{\left(\frac{z}{e} + \frac{-z}{e^{-z}}\right)^2}$$
$$= 1 - (g(z))^2$$

ACTIVATION FUNCTIONS (SUMMARY)

RELU

$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & z < 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z \geq 0 \\ 0; & z < 0 \end{cases}$$

L-RELU

$$g(z) = \begin{cases} z; & z \geq 0 \\ \alpha z; & z < 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z \geq 0 \\ \alpha; & z < 0 \end{cases}$$

SIGMOID

$$g(z) = \frac{1}{1 + e^{-z}}$$

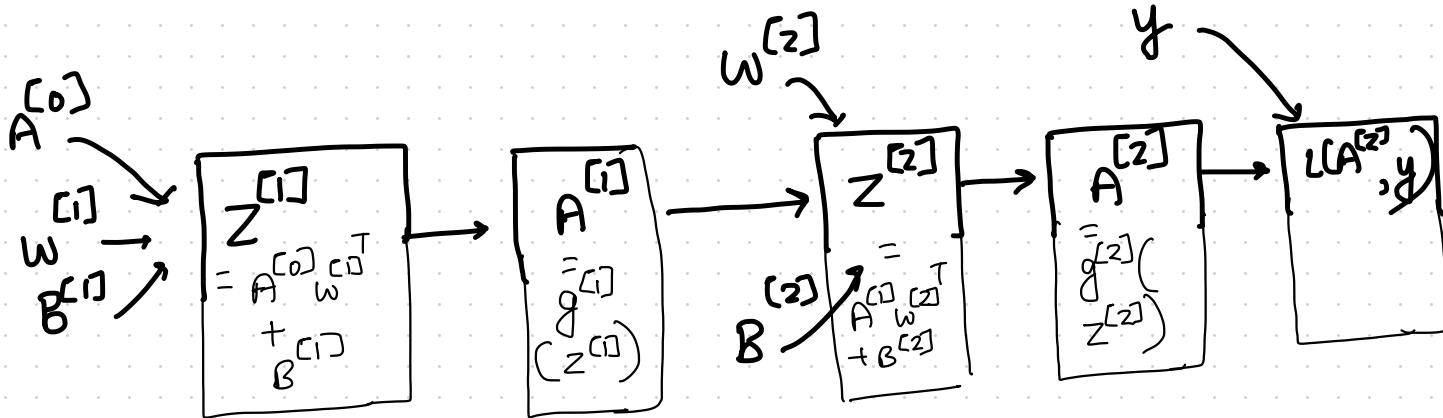
$$g'(z) = g(z) * (1 - g(z))$$

TANH

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$g'(z) = 1 - (g(z))^2$$

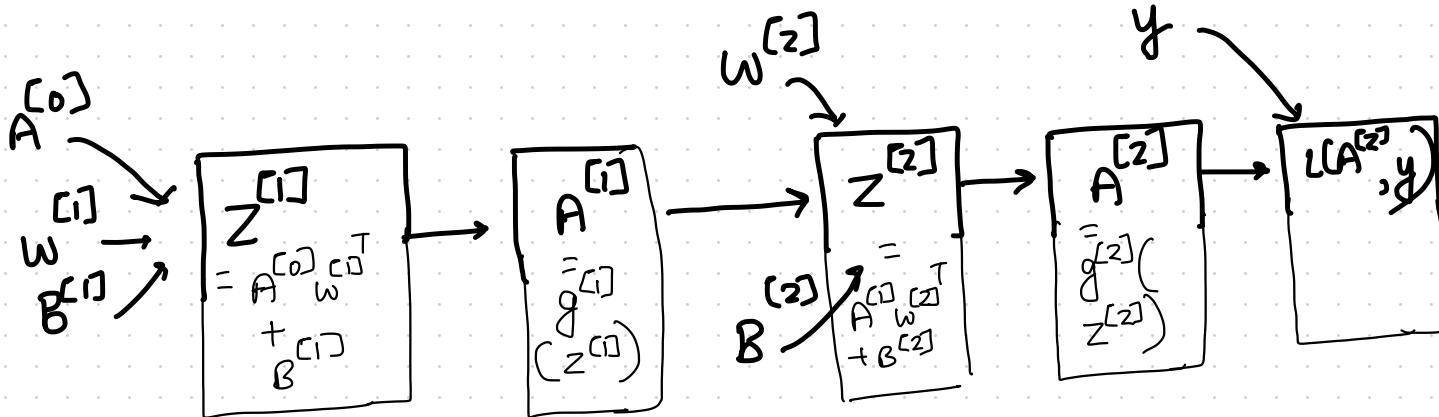
COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$L(A^{[2]}, y) = \sum_{i=1}^N -y_{(i)} \log A^{[2]}_{(i)} - (1 - y_{(i)}) \log (1 - A^{[2]}_{(i)})$$

WRITE IN VECTOR FORM

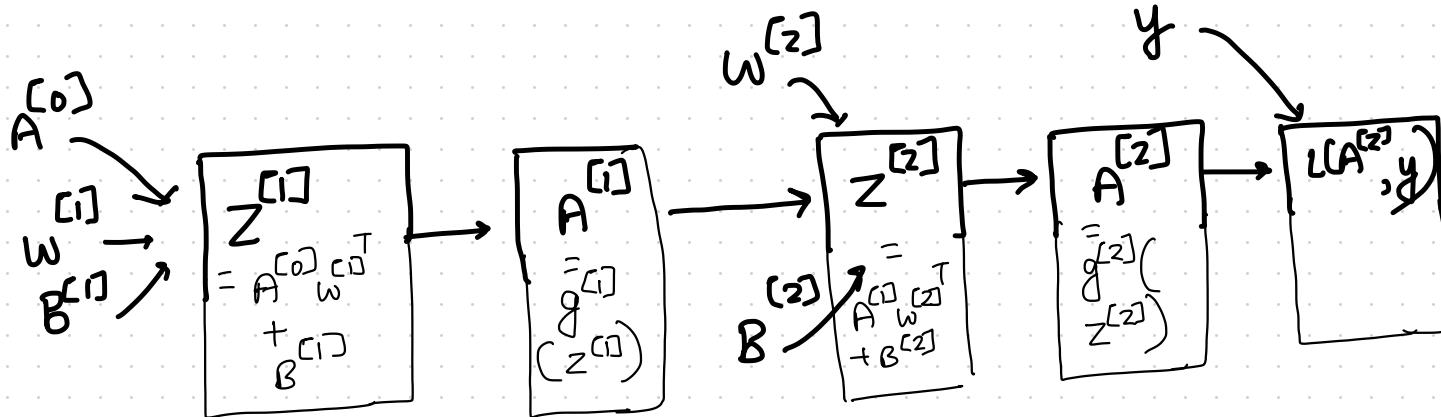
COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\begin{aligned}
 L(A^{[2]}, y) &= \sum_{i=1}^N -y_{(i)} \log A_{(i)}^{[2]} - (1 - y_{(i)}) \log (1 - A_{(i)}^{[2]}) \\
 &= -y^T \log(A^{[2]}) - (1 - y)^T \log(1 - A^{[2]}) \quad [N \times 1]
 \end{aligned}$$

APPLIED ELEMENT-WISE

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\begin{aligned}
 L(A^{[2]}, y) &= \sum_{i=1}^N -y_{(i)} \log A_{(i)}^{[2]} - (1 - y_{(i)}) \log (1 - A_{(i)}^{[2]}) \\
 &= -y^T \log(A^{[2]}) - (1 - y)^T \log(1 - A^{[2]}) \quad [\vdots]_{N \times 1}
 \end{aligned}$$

APPLIED ELEMENT-WISE

$$\frac{\partial L(A^{[2]}, y)}{\partial A^{[2]}} = ?$$

$$\text{let } q_f = y^T \log(A^{[2]})$$

$$q_f = y_{(1)} \log A_{(1)}^{[2]} + \dots + y_{(N)} \log A_{(N)}^{[2]}$$

$$\text{let } q = y^T \log(A^{[2]})$$

$$q = y_{(1)} \log A_{(1)}^{[2]} + \dots + y_{(N)} \log A_{(N)}^{[2]}$$

$$\frac{\partial q}{\partial A^{[2]}} = \begin{bmatrix} \frac{\partial}{\partial A_{(1)}^{[2]}} (y_{(1)} \log A_{(1)}^{[2]} + \dots) \\ \vdots \\ \frac{\partial}{\partial A_{(N)}^{[2]}} (y_{(1)} \dots \dots \dots) \end{bmatrix} = \begin{bmatrix} \frac{y_{(1)}}{A_{(1)}^{[2]}} \\ \vdots \\ \frac{y_{(N)}}{A_{(N)}^{[2]}} \end{bmatrix}$$

$$\frac{\partial q}{\partial A^{[2]}}_{Nx1} = y_{Nx1} \odot A_{Nx1}^{[2]}$$

Element-wise division

$$\text{Let } \alpha = (1 - y)^T \log (1 - A^{[2]})$$

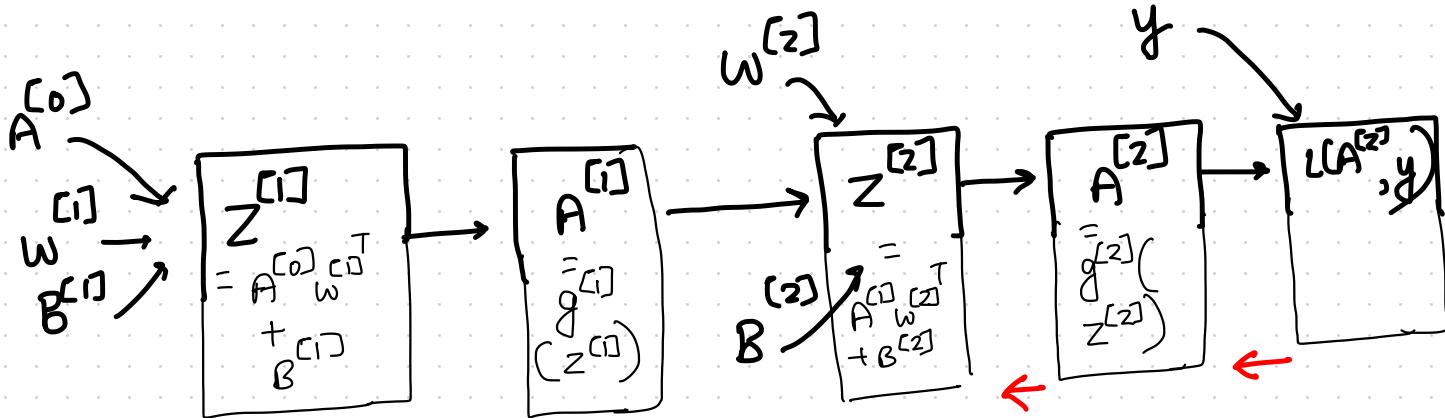
$$\alpha = (1 - y_{(1)}) \log (1 - A_{(1)}^{[2]}) + \dots \dots$$

$$\frac{\partial \alpha}{\partial A^{[2]}} = \left[\begin{array}{c} \frac{\partial}{\partial A_{(1)}^{[2]}} \left\{ (1 - y_{(1)}) \log (1 - A_{(1)}^{[2]}) + \dots \right\} \\ \vdots \\ \frac{\partial}{\partial A_{(N)}^{[2]}} \left\{ (1 - y_{(N)}) \log (1 - A_{(N)}^{[2]}) + \dots \right\} \end{array} \right] = \left[\begin{array}{c} \left(\frac{1 - y_{(1)}}{1 - A_{(1)}^{[2]}} \right) (-1) \\ \vdots \\ \left(\frac{1 - y_{(N)}}{1 - A_{(N)}^{[2]}} \right) (-1) \end{array} \right]$$

$$\frac{\partial \alpha}{\partial A^{[2]}}_{N \times 1} = -(1 - y)_{N \times 1} \odot (1 - A^{[2]})_{N \times 1}$$

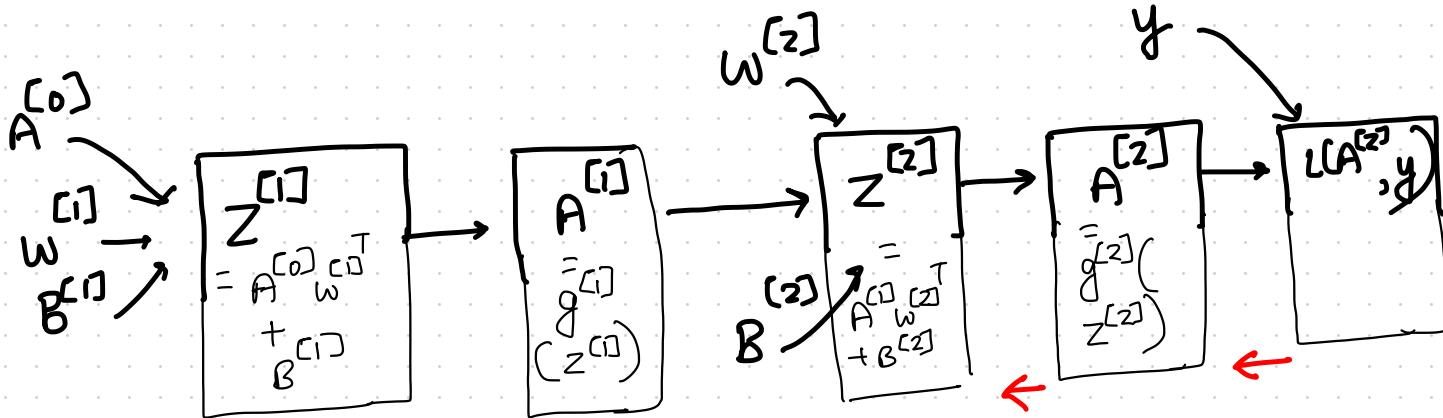
$$\frac{\partial L(A^{[2]}, y)}{\partial A^{[2]}} = -y \odot A^{[2]} + (1-y) \odot (1-A^{[2]})$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial z^{[2]}} = ? \quad \frac{\partial L}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial z^{[2]}} \quad (\text{chain Rule})$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial z^{[2]}} = ? \quad \frac{\partial L}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial z^{[2]}} \quad (\text{Chain Rule})$$

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$\therefore g^{[2]} = \text{sigmoid}$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} \circ A^{[2]} \circ (1 - A^{[2]})$$

$N \times 1$ $N \times 1$ $N \times 1$

Element-wise multiply

$$= (-y \circ A^{[2]} + (1-y) \circ (1-A^{[2]})) \left((A^{[2]} \odot (1-A^{[2]})) \right)$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} A^{[2]} \odot (1 - A^{[2]})$$

$N \times 1$ $N \times 1$

Element-wise

$$= (-y \odot A^{[2]} + (1-y) \odot (1-A^{[2]})) \left(\begin{pmatrix} A^{[2]} & 1 - A^{[2]} \end{pmatrix} \right)$$

$$= -y \odot A^{[2]} \odot A^{[2]} \odot (1 - A^{[2]}) + (1-y) \odot (1 - A^{[2]}) \odot A^{[2]} \odot (1 - A^{[2]})$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} A^{[2]} \odot (1 - A^{[2]})$$

$N \times 1$ $N \times 1$

Element-wise

$$= (-y \odot A^{[2]} + (1-y) \odot (1-A^{[2]})) \left((A^{[2]} \odot (1-A^{[2]})) \right)$$

$$= -y \odot A^{[2]} \odot A^{[2]} \odot (1-A^{[2]}) + (1-y) \odot (1-A^{[2]}) \odot A^{[2]} \odot (1-A^{[2]})$$

$$= -y \odot (1-A^{[2]}) + (1-y) \odot A^{[2]} = -y_{\text{nx1}} + y_{\text{nx1}} \odot A^{[2]}_{\text{nx1}} + A^{[2]}_{\text{nx1}} - y_{\text{nx1}} \odot A^{[2]}_{\text{nx1}}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} A^{[2]} \odot (1 - A^{[2]})$$

$N \times 1$ $N \times 1$

Element-wise

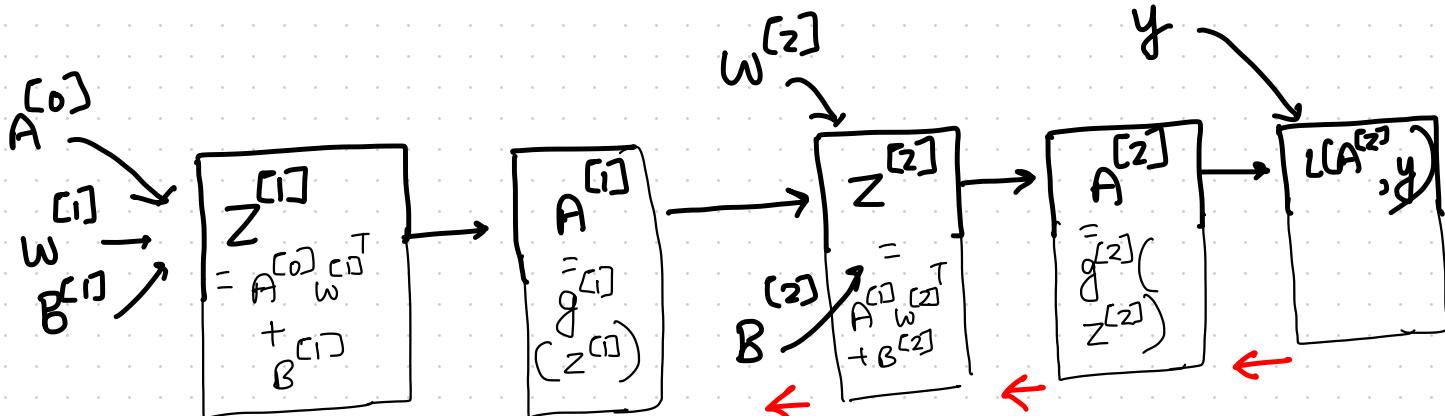
$$= (-y \odot A^{[2]} + (1-y) \odot (1-A^{[2]})) \left((A^{[2]} \odot (1-A^{[2]})) \right)$$

$$= -y \odot A^{[2]} \odot A^{[2]} \odot (1-A^{[2]}) + (1-y) \odot (1-A^{[2]}) \odot A^{[2]} \odot (1-A^{[2]})$$

$$= -y \odot (1-A^{[2]}) + (1-y) \odot A^{[2]} = -y \underset{N \times 1}{+} y \odot A^{[2]} \underset{N \times 1}{+} A^{[2]} \underset{N \times 1}{-} y \odot A^{[2]} \underset{N \times 1}{}$$

$\frac{\partial L}{\partial z^{[2]}}$	$= A^{[2]} - y$
---------------------------------------	-----------------

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial w^{[2]}} = \frac{\partial L}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial w^{[2]}}$$

$\in R^{N^{[2]} \times N^{[1]}}$ (Chain Rule)

SAME DIMENSION ($\because L$ is scalar)

A SIDE

- * GRADIENT: VECTOR IN, SCALAR OUT
- * JACOBIAN : VECTOR IN, VECTOR OUT

$$f: \mathbb{R}^N \rightarrow \mathbb{R}^M$$

$$\text{IP: } \mathbb{R}^N$$

$$\text{OP: } \mathbb{R}^M$$

$$y = f(x)$$

Derivative of 'f' at 'x' called Jacobian is $m \times n$ matrix

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_N} \end{pmatrix}_{m \times N}$$

GENERALISED JACOBIAN: TENSOR IN, TENSOR OUT

$$f: \mathbb{R}^{N_1 \times \dots \times N_D x} \rightarrow \mathbb{R}^{M_1 \times \dots \times M_D y}$$

I/P: D_x - dimensional tensor of shape $N_1 \times \dots \times N_D x$

O/P: D_y - dimensional tensor of shape $M_1 \times \dots \times M_D y$

$$y = f(x)$$

Then; $\frac{\partial y}{\partial x} = \text{Gen. Jacobian} = (M_1 \times \dots \times M_D y) \times (N_1 \times \dots \times N_D x)$
shape

BACK PROP. WITH TENSORS

Let $G_1 = \alpha \beta$

$\alpha: N \times D$

$\beta: D \times M$

$G_1: N \times M$

BACK PROP. WITH TENSORS

Let $G_1 = \alpha \beta$

$$\alpha : N \times D$$

$$\beta : D \times M$$

$$G_1 : N \times M$$

→ We are given loss L as function of G_1

→ we also know $\frac{\partial L}{\partial G_1}$

BACK PROP. WITH TENSORS

Let $G_1 = \alpha \beta$

$$\alpha: N \times D$$

$$\beta: D \times M$$

$$G_1: N \times M$$

→ We are given Loss L as function of G_1

→ we also know $\frac{\partial L}{\partial G_1}$

→ Q: $\frac{\partial L}{\partial \alpha} = ?$; $\frac{\partial L}{\partial \beta} = ?$

BACK PROP. WITH TENSORS

Let $G_1 = \alpha \beta$

$\alpha: N \times D$

$\beta: D \times M$

$G_1: N \times M$

Let's choose: $N=1; D=2; M=3$

BACK PROP. WITH TENSORS

Let $G_1 = \alpha \beta$

$$\alpha : N \times D$$

$$\beta : D \times M$$

$$G_1 : N \times M$$

Let's choose: $N=1; D=2; M=3$

$$G_{1 \times 3} = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} \end{pmatrix}$$

$$\alpha_{1 \times 2} = (\alpha_{1,1} \quad \alpha_{1,2})$$

$$\beta_{2 \times 3} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix}$$

BACK PROP. WITH TENSORS

$$G_{1 \times 3} = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} \end{pmatrix}$$

$$\alpha_{1 \times 2} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \end{pmatrix}$$

$$\beta_{2 \times 3} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix}$$

$$\alpha\beta = \left[\alpha_{1,1} \beta_{1,1} + \alpha_{1,2} \beta_{2,1} ; \alpha_{1,1} \beta_{1,2} + \alpha_{1,2} \beta_{2,2} ; \alpha_{1,1} \beta_{1,3} + \alpha_{1,2} \beta_{2,3} \right]$$

BACK PROP. WITH TENSORS

$$G_{1 \times 3} = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} \end{pmatrix}$$

$$\alpha_{1 \times 2} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \end{pmatrix}$$

$$\beta_{2 \times 3} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix}$$

$$\alpha\beta = \left[\alpha_{1,1} \beta_{1,1} + \alpha_{1,2} \beta_{2,1} ; \alpha_{1,1} \beta_{1,2} + \alpha_{1,2} \beta_{2,2} ; \alpha_{1,1} \beta_{1,3} + \alpha_{1,2} \beta_{2,3} \right]$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = \left[\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3} \right]$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = \left[\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3} \right]$$

BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}] : \text{Shape} = \begin{matrix} \text{Shape of } \alpha\beta \\ \alpha_{1,1} \end{matrix} \times \text{Shape of } \beta_{1,1}$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}] : \text{Shape} = (N \times M) \times 1$$

BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}] : \text{Shape} = \begin{matrix} \text{Shape of } \alpha\beta \times \text{Shape of} \\ \alpha_{1,1} \end{matrix}$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}] : \text{Shape} = (N \times M) \times 1$$

Generalised
Tensor shape

BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}]$$

$\frac{\partial}{\partial \alpha_{1,1}}$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}]$$

$\frac{\partial}{\partial \alpha_{1,2}}$

We know $\frac{\partial L}{\partial g} = [dg_{1,1} \quad dg_{1,2} \quad dg_{1,3}]$

Generalised
shape =
 $1 \times (N \times M)$
↳ why?

BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}]$$

$\frac{\partial}{\partial \alpha_{1,1}}$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}]$$

$\frac{\partial}{\partial \alpha_{1,2}}$

We know $\frac{\partial L}{\partial g} = [dg_{1,1} \quad dg_{1,2} \quad dg_{1,3}]$

Generalised
shape =

$1 \times (N \times M)$



$\therefore L$ is a
scalar

BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}]$$

$\partial \alpha_{1,1}$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}]$$

$\partial \alpha_{1,2}$

We know $\frac{\partial L}{\partial g} = [dg_{1,1} \quad dg_{1,2} \quad dg_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[\begin{array}{cc} \frac{\partial L}{\partial \alpha_{1,1}} & \frac{\partial L}{\partial \alpha_{1,2}} \end{array} \right]$$

BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}] = \frac{\partial G}{\partial \alpha_{1,1}}$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}] = \frac{\partial G}{\partial \alpha_{1,2}}$$

We know $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[\frac{\partial L}{\partial \alpha_{1,1}} \quad \frac{\partial L}{\partial \alpha_{1,2}} \right]$$

$$\frac{\partial L}{\partial \alpha_{1,1}} = \frac{\partial L}{\partial G} \cdot \frac{\partial G}{\partial \alpha_{1,1}}$$

BACK PROP. WITH TENSORS

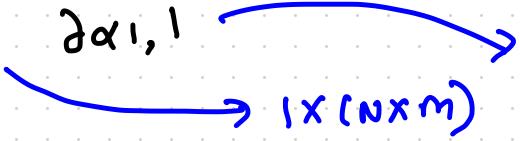
$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}] = \frac{\partial G}{\partial \alpha_{1,1}}$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}] = \frac{\partial G}{\partial \alpha_{1,2}}$$

We know $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[\frac{\partial L}{\partial \alpha_{1,1}} \quad \frac{\partial L}{\partial \alpha_{1,2}} \right]$$

$$\frac{\partial L}{\partial \alpha_{1,1}} = \frac{\partial L}{\partial G} \cdot \frac{\partial G}{\partial \alpha_{1,1}}$$


 $1 \times (N \times M)$ $(N \times M) \times 1$

BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = [\beta_{1,1}; \beta_{1,2}; \beta_{1,3}] = \frac{\partial G}{\partial \alpha_{1,1}}$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = [\beta_{2,1}; \beta_{2,2}; \beta_{2,3}] = \frac{\partial G}{\partial \alpha_{1,2}}$$

We know $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[\frac{\partial L}{\partial \alpha_{1,1}} \quad \frac{\partial L}{\partial \alpha_{1,2}} \right]$$

$$\frac{\partial L}{\partial \alpha_{1,1}} = \frac{\partial L}{\partial G} \cdot \frac{\partial G}{\partial \alpha_{1,1}} = [dG_{1,1} \beta_{1,1} + dG_{1,2} \beta_{1,2} + dG_{1,3} \beta_{1,3}]$$

$(1 \times (n \times m)) \times ((n \times m) \times 1) \rightarrow$ Dot product of two.

BACK PROP. WITH TENSORS

$$\frac{\partial L}{\partial \alpha} = \left[\frac{\partial L}{\partial \alpha_{1,1}} ; \frac{\partial L}{\partial \alpha_{1,2}} \right] = \begin{pmatrix} \frac{\partial L}{\partial G_1} \\ \vdots \\ \frac{\partial L}{\partial G_m} \end{pmatrix}$$

$1 \times (N \times D)$ $1 \times (N \times m)$ $m \times D$

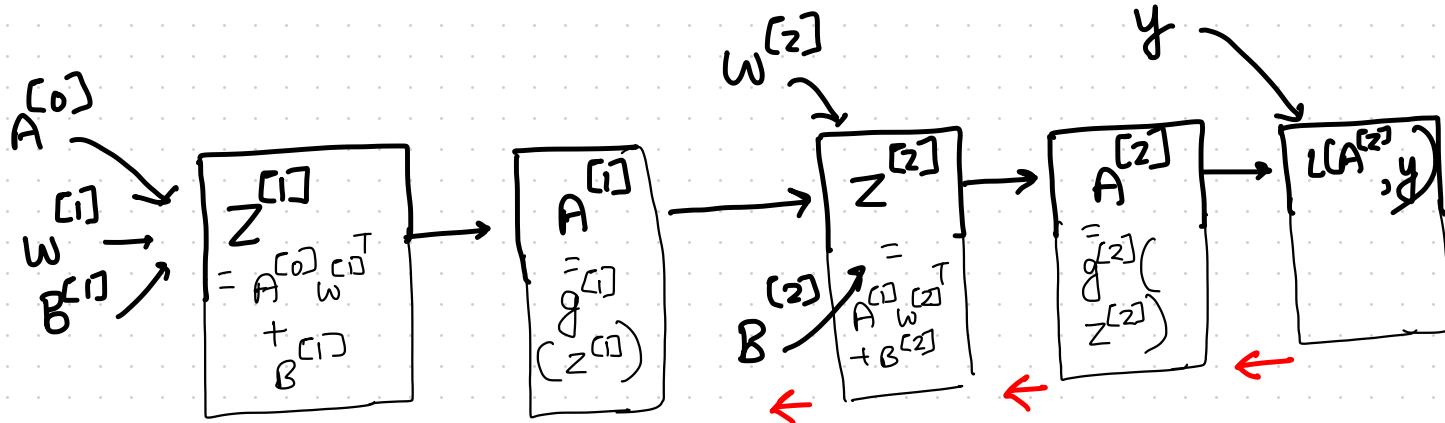
BACK PROP. WITH TENSORS

$$\frac{\partial L}{\partial \alpha} = \left[\frac{\partial L}{\partial \alpha_{1,1}} ; \frac{\partial L}{\partial \alpha_{1,2}} \right] = \begin{pmatrix} \frac{\partial L}{\partial G} \end{pmatrix}^T \quad \begin{matrix} 1 \times (N \times D) \\ 1 \times (N \times m) \end{matrix} \quad \begin{matrix} G^T \\ m \times D \end{matrix}$$

Similarly;

$$\frac{\partial L}{\partial \beta} = \alpha^T \quad \begin{matrix} \alpha^T \\ D \times N \end{matrix} \quad \frac{\partial L}{\partial G} \quad \begin{matrix} 1 \times (N \times m) \\ 1 \times (D \times m) \end{matrix}$$

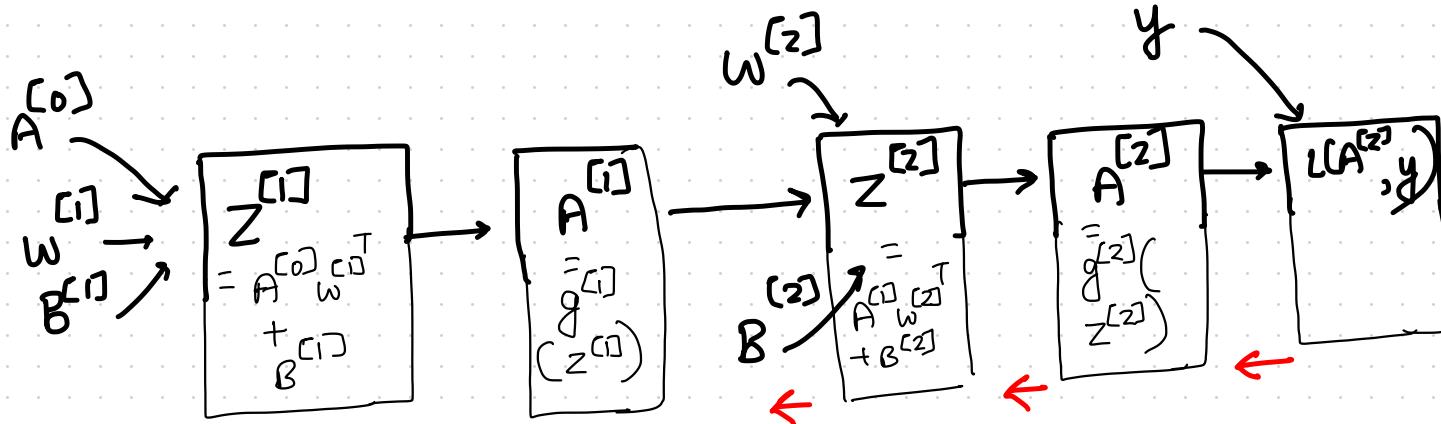
COMPUTATION GRAPH (FOR XOR EXAMPLE)



Equivalence from Aside

$$\begin{aligned}
 G &\leftrightarrow Z^{[2]} \\
 \alpha &\leftrightarrow A^{[1]} \\
 \beta &\leftrightarrow w^{[2] T}
 \end{aligned}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



Equivalence from Aside

$$G \leftrightarrow Z^{[2]}$$

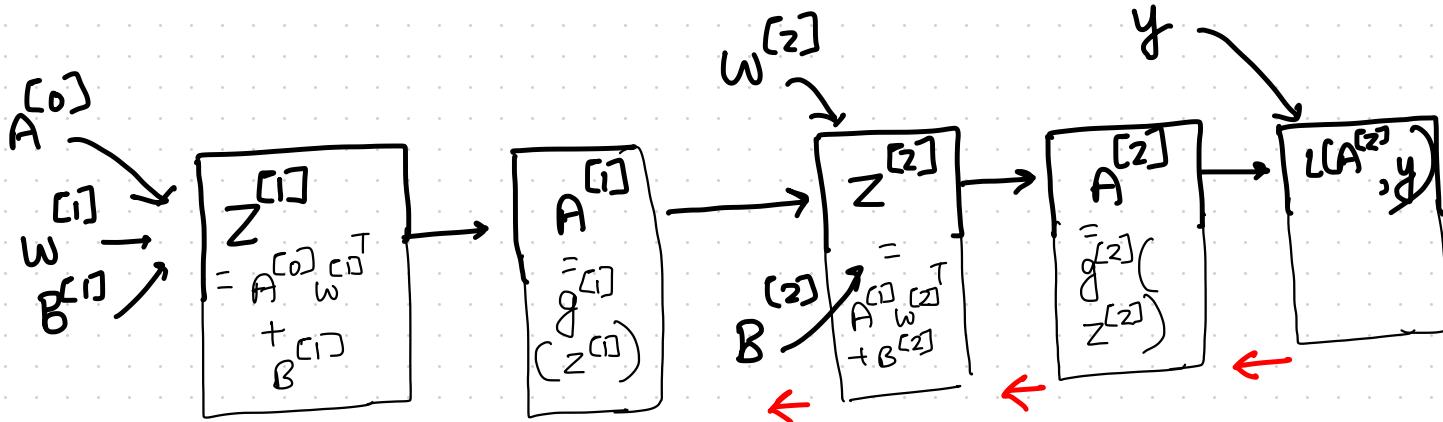
$$\alpha \leftrightarrow A^{[1]}$$

$$\beta \leftrightarrow w^{[2] T}$$

$$\frac{\partial L}{\partial w^{[2] T}} = A^{[1]} \frac{\partial L}{\partial Z^{[2]}}$$

$$= A^{[1] T} \left(A^{[2]} - y \right)$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

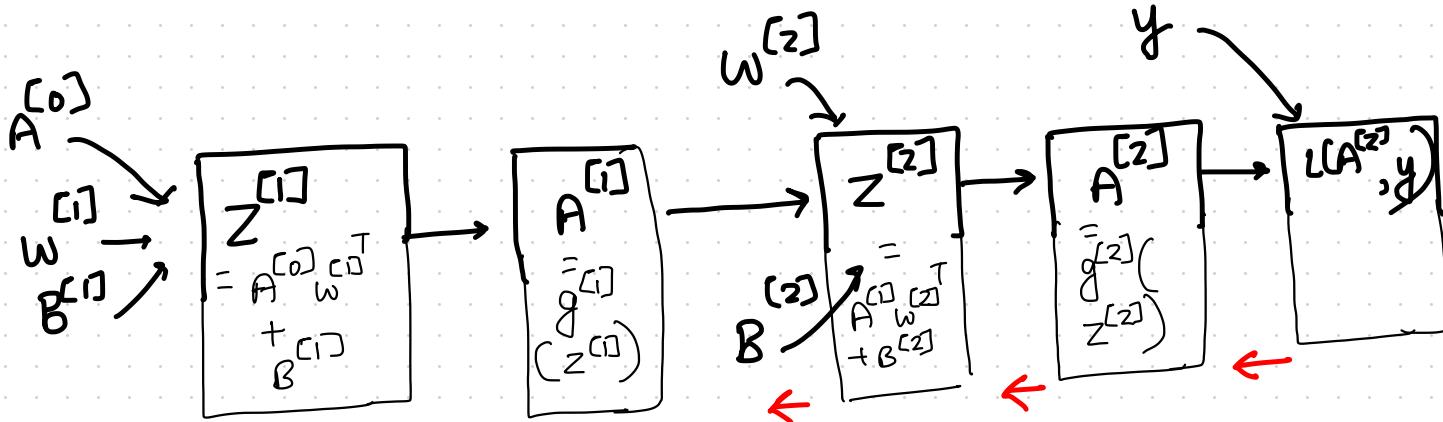


$$\frac{\partial L}{\partial w^{[2]T}} = A^{[1]T} (A^{[2]} - y)$$

$$\Rightarrow \frac{\partial L}{\partial w^{[2]}} = (A^{[2]} - y)^T \underset{N \times N}{A^{[1]}} \underset{N \times N^{[1]}}{A^{[1]}}$$

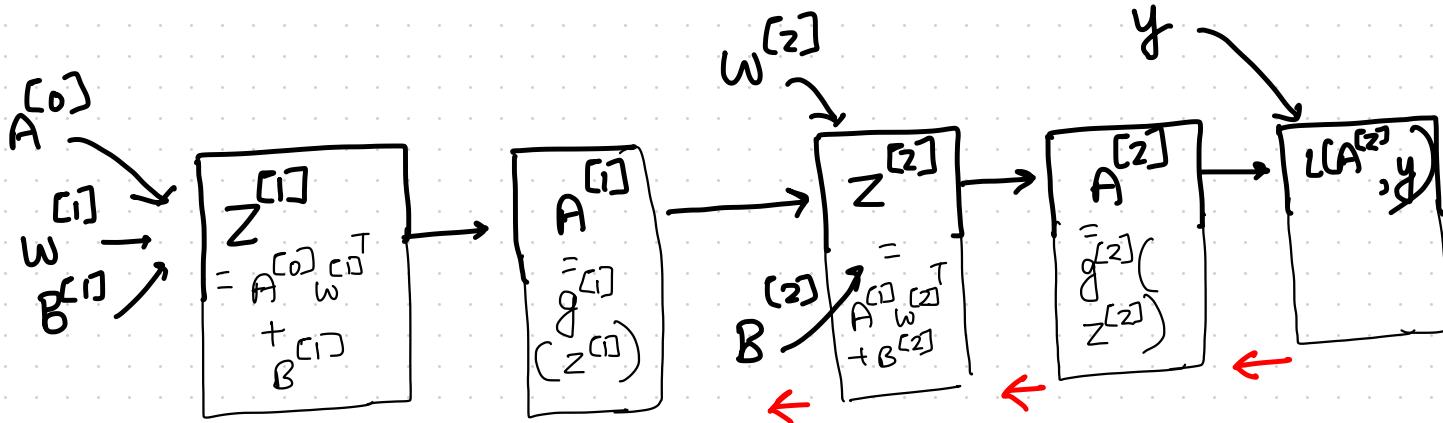
$(N^{[2]} \times N^{[1]})$
 $(N^{[2]} \times N^{[1]})$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial B^{[2]}} = ?$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



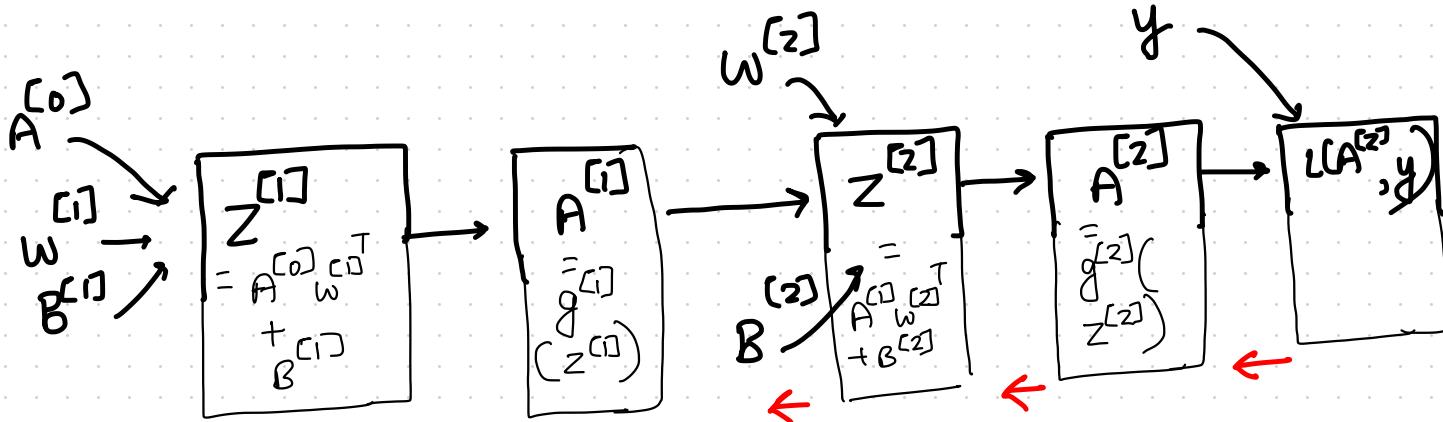
$$\frac{\partial L}{\partial B^{[2]}} = ?$$

$$T$$

$$\text{Let } C = b^{[2]} T$$

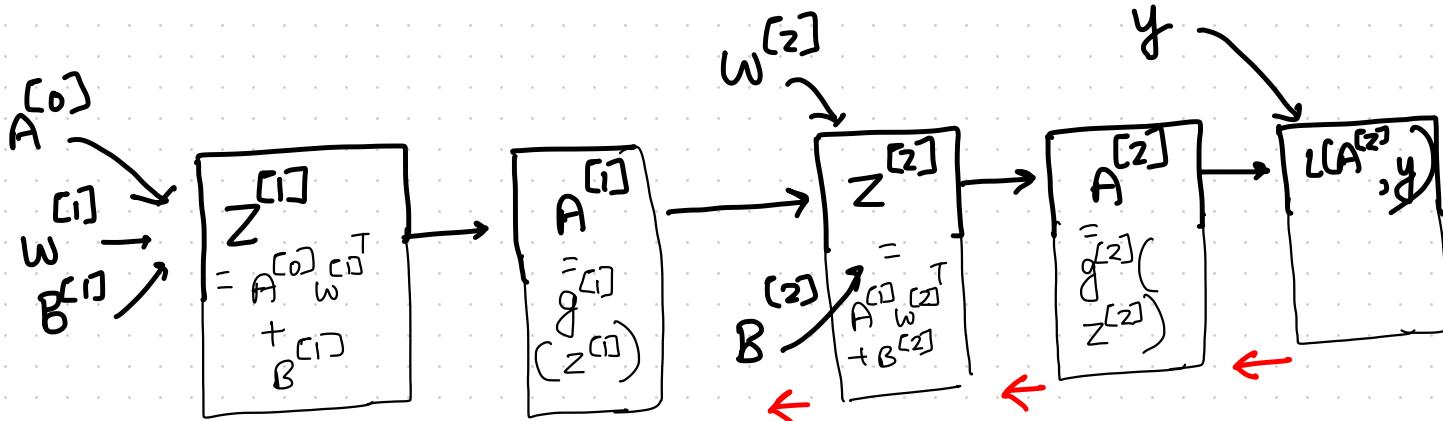
$$B^{[2]} = \begin{bmatrix} - & C & - \\ - & C & - \\ \vdots & \ddots & \end{bmatrix}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



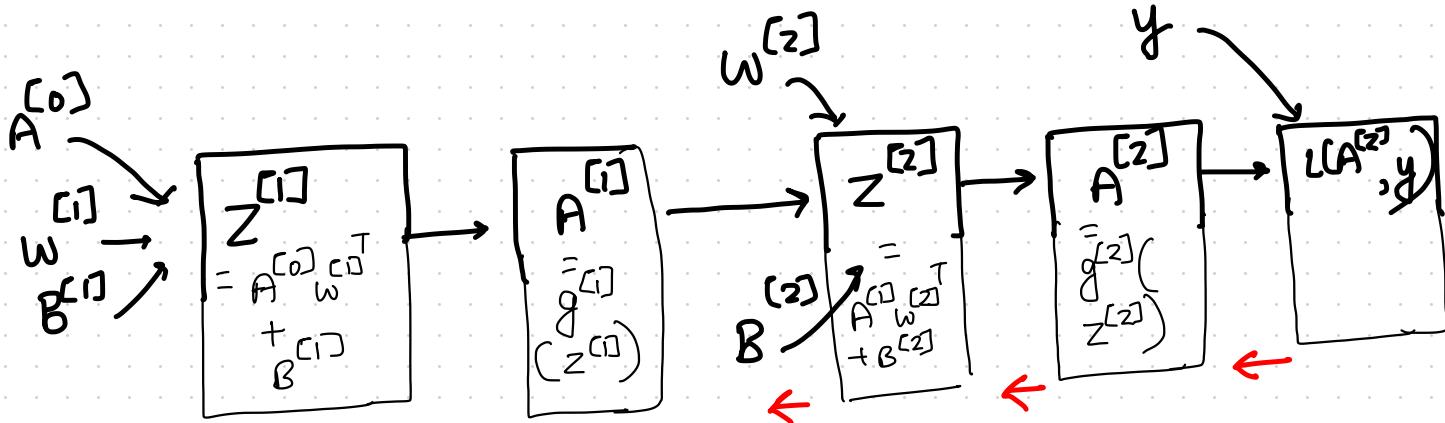
$$\frac{\partial L}{\partial c} = ?$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\begin{aligned} \frac{\partial L}{\partial C} &= \left[\frac{\partial L}{\partial C_1}, \dots, \frac{\partial L}{\partial C_{N^{[2]}}} \right] \\ &= \left[\frac{\partial L}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial C_1}, \dots \right] \end{aligned}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

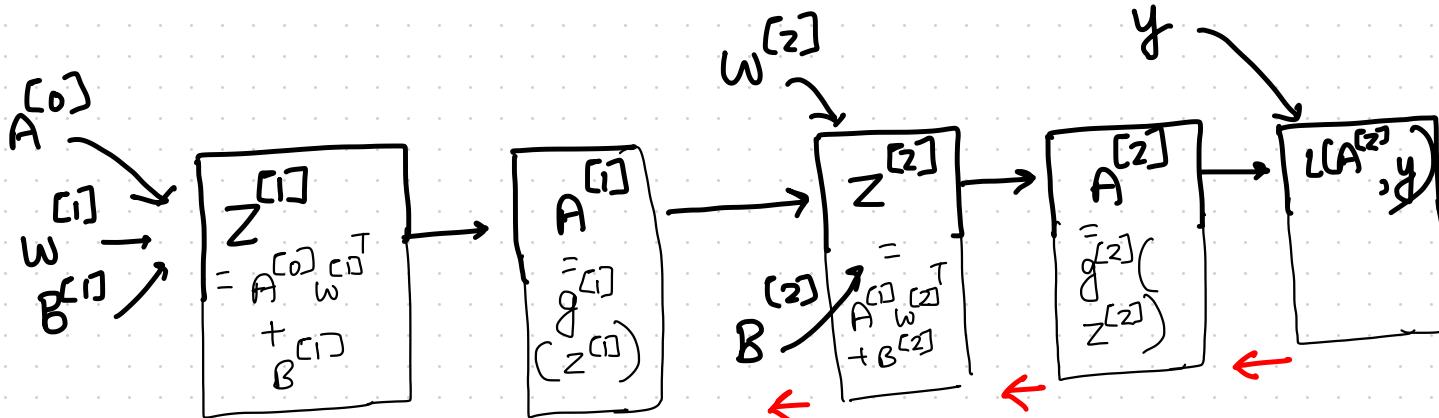


$$\frac{\partial L}{\partial c_1} = \frac{\partial L}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial c_1}$$

$$z^{[2]} = \begin{bmatrix} z_{1,1}^{[2]} & \dots & z_{1,N^{[2]}}^{[2]} \\ \vdots & & \vdots \\ z_{N,1}^{[2]} & \dots & z_{N,N^{[2]}}^{[2]} \end{bmatrix} = A^{[2]} w^{[2]}{}^T + \begin{bmatrix} c_1 & c_2 & \dots & c_N^{[2]} \\ c_1 & c_2 & \dots & c_N^{[2]} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

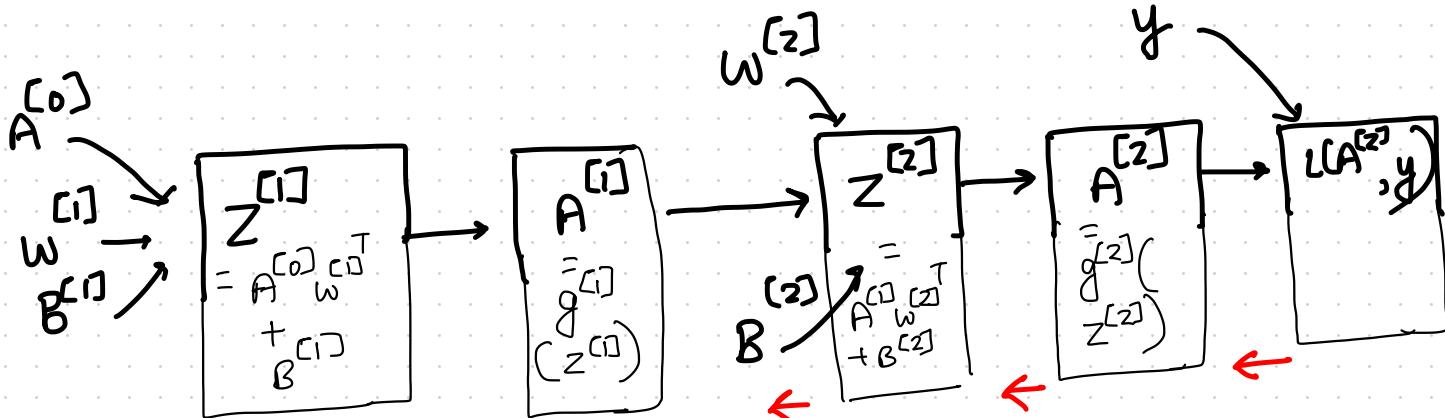
$N \times N^{[2]}$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial Z^{[2]}}{\partial c_1} = \begin{bmatrix} \frac{\partial c_1}{\partial c_1} & \frac{\partial c_2}{\partial c_1} & \dots \\ \vdots & \ddots & \dots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & \ddots & \ddots & \ddots & \vdots \\ 1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \end{bmatrix} \quad (N \times N^{[2]}) \times (1)$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

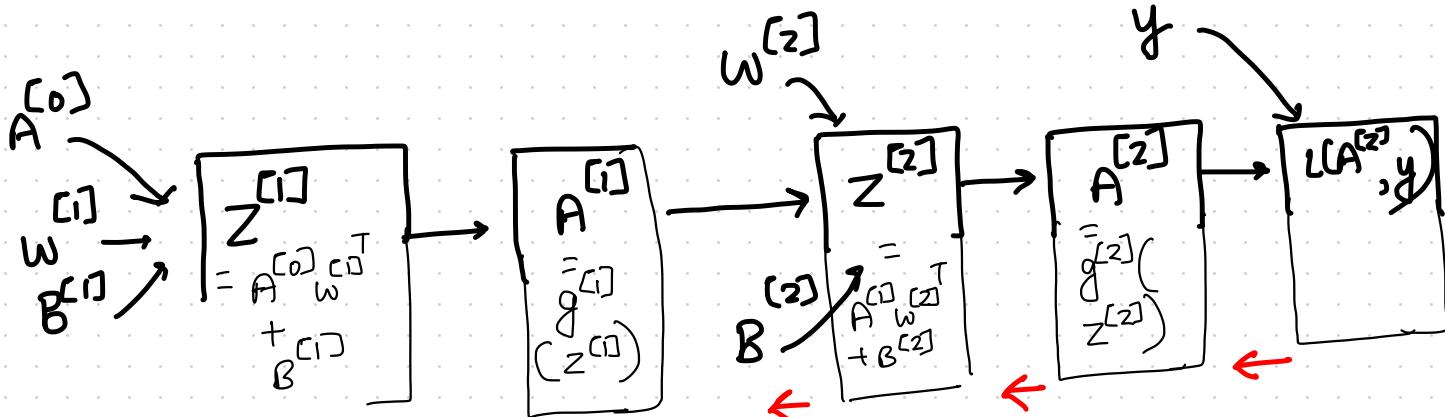


$$\frac{\partial Z}{\partial c_i} = \begin{bmatrix} \frac{\partial c_1}{\partial c_1} & \frac{\partial c_2}{\partial c_1} & \dots \\ \vdots & \ddots & \dots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & & & \\ 1 & - & - & - & - \end{bmatrix} \quad (N \times N^{[2]}) \times (1)$$

$$\frac{\partial L}{\partial c_i} = \frac{\partial L}{\partial Z^{[2]}} \quad \frac{\partial Z^{[2]}}{\partial c_i}$$

$(N \times N^{[2]}) \times 1$ $(N \times N^{[2]}) \times 1$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\text{Let } \frac{\partial L}{\partial z^{[2]}} = \begin{bmatrix} dz_{1,1}^{[2]}, \dots, dz_{1,N^{[2]}}^{[2]} \\ dz_{N,1}^{[2]}, \dots, dz_{N,N^{[2]}}^{[2]} \end{bmatrix}$$

Then

$$\frac{\partial L}{\partial c_1} = \sum_{i=1}^N dz_{i,1}^{[2]}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

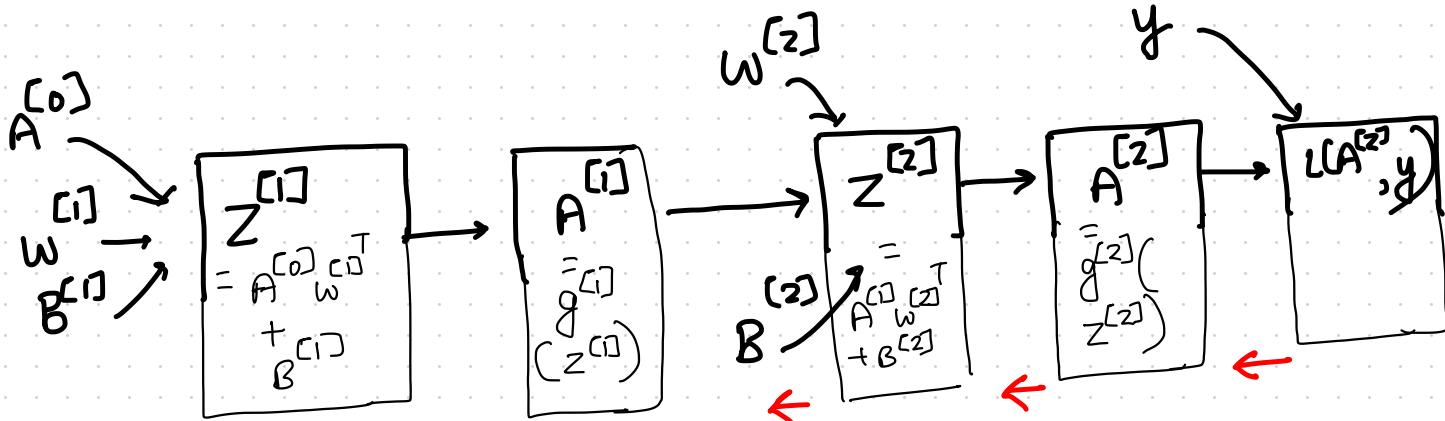
$$\frac{\partial L}{\partial c_1} = \sum_{i=1}^N dZ_i^{[2]}, 1$$

$$\Rightarrow \frac{\partial L}{\partial c} = \left[\sum_{i=1}^N dZ_{i,1}^{[2]}, \sum_{i=1}^N dZ_{i,2}^{[2]}, \dots, \sum_{i=1}^N dZ_{i,N}^{[2]} \right]$$

$$\Rightarrow \frac{\partial L}{\partial b} = \left(\frac{\partial L}{\partial c} \right)^T = \begin{bmatrix} \sum_{i=1}^N dZ_{i,1}^{[2]} \\ \vdots \\ \sum_{i=1}^N dZ_{i,N}^{[2]} \end{bmatrix}$$

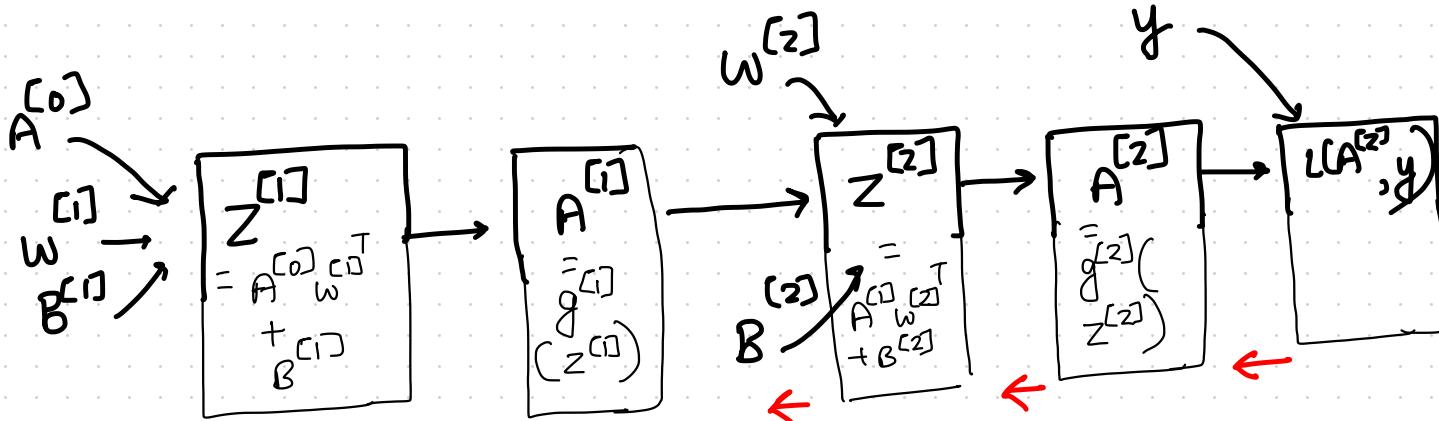
$N^{[2]} \times 1$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial A^{[0]}} = \frac{\partial L}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial A^{[1]}} = \frac{\partial L}{\partial Z^{[2]}} \left(w^{[2]} \right)^T = \frac{\partial L}{\partial Z^{[2]}} w^{[2]}$$

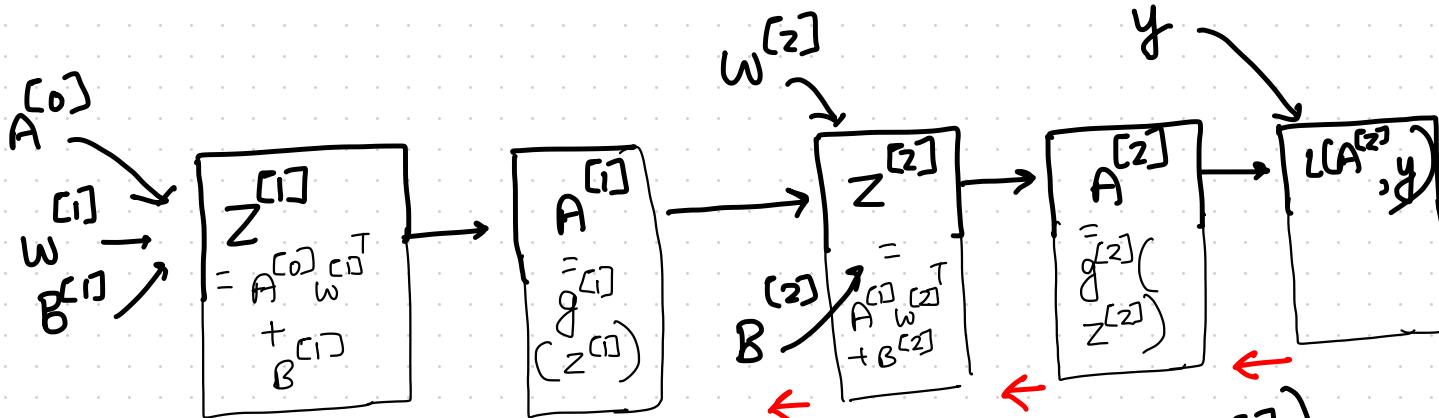
COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial A^{[1]}} = \frac{\partial L}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial A^{[1]}} = \frac{\partial L}{\partial Z^{[2]}} (w^{[2]})^T = \frac{\partial L}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial A^{[2]}} w^{[2]}$$

$1 \times (N \times N^{[1]})$ $1 \times (N \times N^{[2]})$ $N^{[2]} \times N^{[1]}$

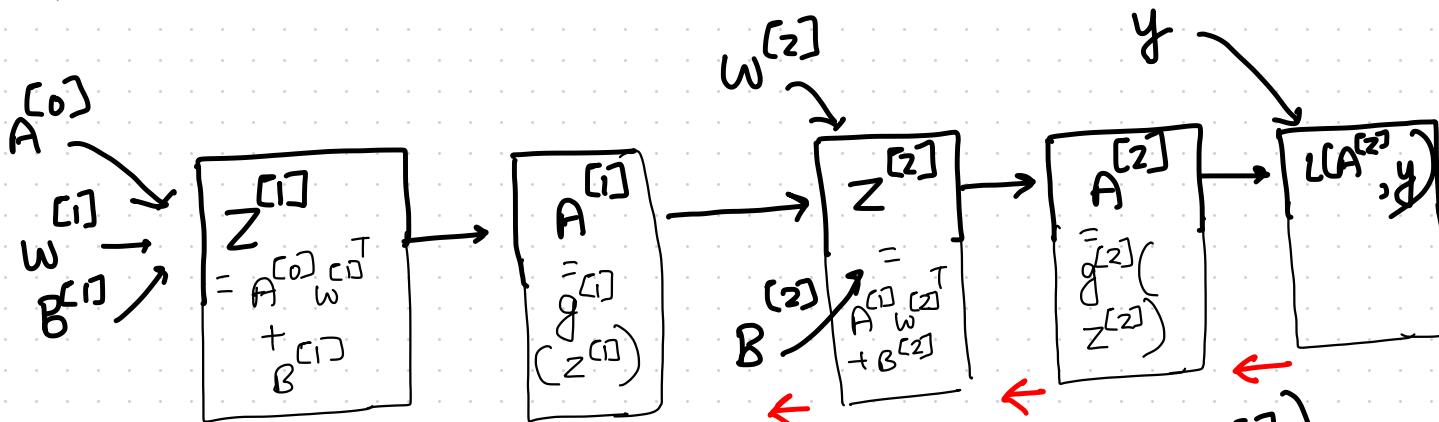
COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial Z^{[2]}} = \frac{\partial L}{\partial A^{[1]}} \frac{\partial A^{[1]}}{\partial Z^{[1]}} = \left(\frac{\partial L}{\partial Z^{[2]}} w^{[2]} \right) \left(\frac{\partial A^{[1]}}{\partial Z^{[1]}} \right)$$

(1x(NxN^[1]))
 1x(NxN^[2]) x (N^[2] x N^[1])
 (NxN^[1]) x (NxN^[2])

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial Z^{[1]}} = \frac{\partial L}{\partial A^{[1]}} \frac{\partial A^{[1]}}{\partial Z^{[1]}} = \left(\frac{\partial L}{\partial Z^{[2]}} w^{[2]T} \right) \left(\frac{\partial A^{[1]}}{\partial Z^{[1]}} \right)$$

$(N \times N^{[1]})$

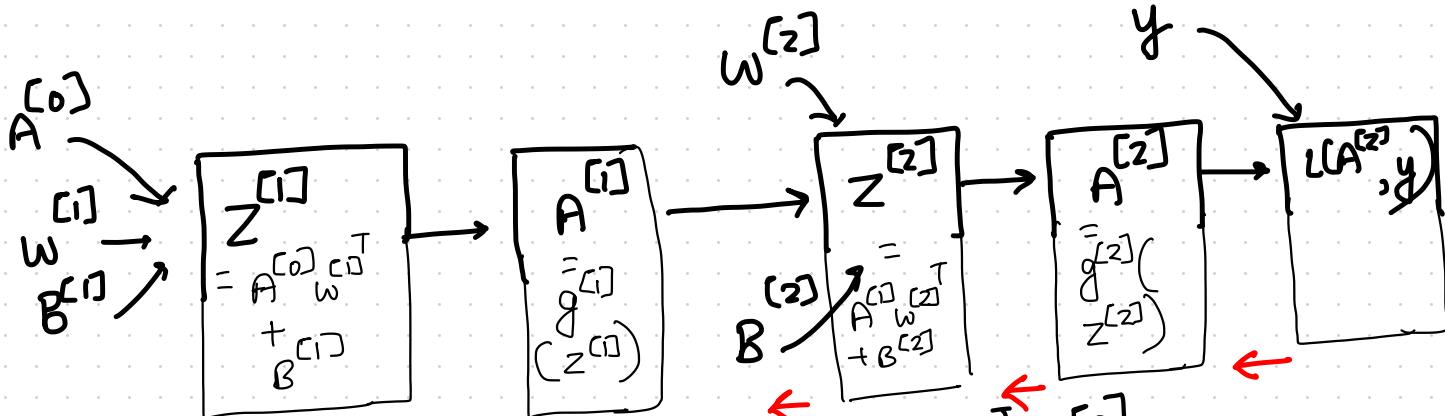
$1 \times (N \times N^{[2]}) \times (N^{[2]} \times N^{[1]})$

$(N \times N^{[1]})$

$$= \frac{\partial L}{\partial Z^{[1]}} w^{[2]T} \circ g^{[1]'}(Z^{[1]})$$

Element wise
or dot product of Jacobian

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial w^{[1]}} = A^{[0]T} \frac{\partial L}{\partial z^{[1]}} \Rightarrow \frac{\partial L}{\partial w^{[1]}} = \left(\frac{\partial L}{\partial z^{[1]}} \right)^T A^{[0]}$$

$$\frac{\partial L}{\partial b^{[1]}} = \begin{bmatrix} \sum_{i=1}^N dz_{i,1} \\ \vdots \\ \sum_{i=1}^N dz_{i,N}^{[2]} \end{bmatrix}_{N^{[2]} \times 1}$$